The Impact of Heavy-tailed Error Distributions on Partially Nested Randomized Controlled Trials Models

A DISSERTATION SUBMITTED TO THE FACULTY OF THE UNIVERSITY OF MINNESOTA

BY

Mario Raul Moreno

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Michael R. Harwell.

September, 2017

Acknowledgments

I am immensely grateful to God. He provided me with the courage and determination to finish this dissertation. I am eternally grateful to my lovely wife, Mariela. She emotionally supported me throughout the hardest times of this endeavor. I could not have done this without her. I am thankful to my children, Rebeca and Xavier. I took time belonging to them to pursue my Ph.D. I would also like to acknowledge my parents for their support. Thank you, Mom and Dad. I want to also thank to my brothers, Boris and Marvin, who always had supportive words for me.

I would like to thank my adviser, Dr. Harwell, for his thoughtful mentoring and commitment to my training and success as a scholar. He pushed me to be a better educational researcher and statistician. I really appreciate all the time he dedicated to my dissertation. I am grateful to Dr. Rodriguez for being my friend and sharing his knowledge with me. I also extend my thanks to Dr. Wang and Dr. Davenport, members of my dissertation committee, for their valuable feedback throughout the various iterations of my dissertation. Additionally, I want to thank Dr. Davison, Dr. Zieffler, Dr. Lawrenz, and Dr. delMas. I received help from them when needed.

Thank you to my friends Amy, Mao, Kyle, Jose, Kory, Martin, Yadira, Nicola, Elizabeth, and Anelise for all those memorable moments we lived together in the program. I am immensely grateful to Ethan Brown for his support in learning R and writing the R script for this dissertation.

Finally, I want to thank Kim Strain and Patricia Eliason for providing me with advice and tools to improve my writing.

Dedication

I dedicate this work to God, my wife (Mariela), my children (Rebeca and Xavier), my parents (Raúl and Guiller), and my brothers (Boris and Marvin).

Abstract

The Partially Nested Randomized Controlled Trial (PNRCT) model can be used when the subjects within the treatment group are clustered in groups and the subjects in the control group remained unclustered. Because this model has been proposed recently, few studies can be found in the literature. The few existing studies have focused on validating, by using Monte Carlo techniques, the efficiency of the PNRCT model compared with other competing models. These studies have been conducted under the assumption that all model assumptions are achieved. However, when fitting regression models, the model assumptions may be a problem. One of these assumptions is the normality of error distribution.

The literature suggests that real world data hardly ever present a normal distribution. When using data from applied settings, the error distribution from regression models can produce any distributional form. When this happens, the model assumption of normality may be violated, affecting the quality of parameter estimates. This is relevant for the validity of any regression model, because if the error distribution substantially deviates from normality, the model parameter estimates and their standard errors may be seriously affected. Despite the relevance of normality error distribution, very little attention has been given to it in partially nested models.

This study assessed the effect of violating the assumption of normality at level 2 in the PNRCT model on the sensitivity of parameter estimates (fixed effects), Type I error rate, and power in the "pure" PNRCT model and in a PNRCT model adjusted by one covariate at level 1. To achieve this goal several conditions such as different number

of clusters, different cluster size, and different intra-class correlation (ICC) levels were examined. The results showed robust estimation of the PNRCT model. The fixed effects were unbiased and were more accurately estimated when the number of clusters and subjects increased. The same pattern was found as the ICC increased. The PNRCT model diminishes power up to certain point. This condition is exacerbated in the presence of non-normal distributions. However, as in other studies, power was positively impacted as the number of clusters and number of subjects increased. Finally, the Type I error rate did not substantially deviate from the nominal Type I error rate even for non-normal distributions.

List of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Random sampling and random assignment are two important concepts in educational research. Random sampling is the process of selecting subjects or units from a population in a random fashion. This implies that every possible sample could be selected with a predetermined probability of being selected. In addition, selecting a simple-random sample, subjects must have the same likelihood to be chosen from the population (Lohr, 2009). On the other hand, random assignment is the process of assigning subjects or units randomly to two or more conditions, treatment(s) or control(s) (Boruch, De Moya, & Snyder, 2002). These concepts are directly associated with external and internal validity in the research process. Whereas external validity allows researchers to make generalizations about the population from which the sample was drawn, internal validity is usually associated with causality. Random assignment contributes to internal validity and is a powerful tool in education research to impute causality of any tested treatment.

Rubin (1974) was able to demonstrate how a randomization process generates an unbiased estimate of the average treatment effect. Random assignment allocates subjects into two or more groups, which results in a very high chance that the two groups will be statistically equivalent in their characteristics (Boruch et al., 2002).

Statistically equivalent groups means that subjects assigned to the control group will have similar characteristics (observable and non-observable variables) to subjects assigned to the treatment group, and the only difference between them will be the treatment condition (Gertler, Martinez, Premand, Rawlings, & Vermeersch, 2011). Therefore, random assignment allows researchers to estimate causal effects on an outcome because random assignment permits them to estimate the output of a counterfactual group, that is, what the outcome would have been if individuals in the treatment condition had not received any treatment. This arrangement permits two important steps in the research process: 1) to make comparisons between two statistically equivalent groups and 2) to make valid statements about the results (Boruch et al., 2002).

In general, there are three approaches for randomized experimental designs that obviously involve random assignment. These approaches are (a) Individual Randomized Controlled Trials (IRCT), (b) Cluster Randomized Controlled Trials (CRCT), and (c) Partially Nested Randomized Control Trials (PNRCT) (Bauer, Sterba, & Hallfors, 2008; Lohr, Schochet, & Sanders, 2014).

In IRCT, individuals (e.g., students) are randomly assigned to treatment and control conditions. In CRCT, a higher level of hierarchy (e.g., schools/clusters) is assigned to treatment and control conditions and all individuals within a cluster receive the same condition, treatment or control. In PNRCT, students are also randomly assigned to treatment and control conditions. However, in the treatment condition students are allocated in groups (clusters) in order to receive the treatment whereas in the control condition individuals remain ungrouped.

Although researchers report methods for properly analyzing IRCT and CRCT, little research has been done in PNRCT, and only recently have some models been proposed and validated for properly analyzing PNRCT data structures (Bauer et al., 2008; Lee & Thompson, 2005; Lohr et al., 2014; Roberts & Roberts, 2005). These PNRCT models have been validated by performing Monte Carlo studies. The results of these studies suggest that these models should be used instead of alternative models that do not properly take into account the nested data structure, which implies the use of multilevel modeling.

PNRCT models and its extensions have been validated assuming that all model assumptions hold. However, real world data may produce models that violate some of the assumptions. For that reason, it is important to know the model assumptions and the impact of violating these model assumptions by evaluating the performance of the PNRCT model when some assumptions do not hold. Among these model assumptions, the normality assumption at level 2 is the one I am interested in in this research.

One important consideration is that the PNRCT model and the PNRCT data analysis are considered the same in this paper and will be used interchangeably. Additionally, in the PNRCT literature, the terms *treatment* and *control conditions* can be found as *treatment* and *control groups*, or *treatment* and *control arms* (Baldwin et al., 2011; Bauer et al. 2008). From here on, the terms *treatment (control) group*, *treatment (control) arm* and *treatment (control) condition* are used interchangeably.

1.1.The PNRCT Model

In the PNRCT design randomization is done on an individual basis, distributing individuals into treatment and control conditions. Then, individuals in the treatment group receive the intervention in clusters. They may or may not be randomly assigned into these clusters. The same treatment is administered in a group setting so that multiple individuals receive the treatment together. In the control group students remain unclustered (Bauer et al., 2008; Lee & Thompson, 2005; Lohr et al, 2014).

An example of the PNRTC design is a research study that evaluated a computer-based balanced literacy intervention. Savage, Abrami, Hipps, and Deault (2009) randomly assigned children (N= 144) from first grade into one of two experimental groups-- a synthetic phonics method and an analytic phonics method--and one control group with no intervention. The two intervention conditions were computer-based, and students (n_1 = 43 and n_2 = 44) received the interventions four times every week for a period of twelve weeks. Children were taught in groups (clusters) of four. The control group (n = 57) received regular instruction at the same time that the other students were instructed with the intervention methods. This is an example of PNRCT design because of three factors: (a) students were assigned to treatment and control conditions, (b) within the treatment conditions students were assigned to clusters, and (c) students within the control condition remained unclustered.

Another example of the PNRCT design is the study, "Family Group Cognitive–Behavioral Preventive Intervention for Families of Depressed Parents: 18- and 24-Month Outcomes" (Compas et al., 2011). In this study, Compas et al. examined the impact of a

family group cognitive-behavioral (FGCB) preventive intervention on mental health outcomes for families with a history of major depressive disorder (MDD).

The participants (111 families) were randomly assigned to a treatment or control condition. The FGCB intervention was a manualized 12-session program (8 weekly and 4 monthly sessions) while controls received written materials to provide education about the nature of depression, effects of parental depression on families, and signs of depression in children. The treatment condition was comprised of small groups (14 groups) of up to four families in each group (in total 56 families – n_1 = 56-). In contrast, the participants (n_2 = 55) in the control condition were not nested within any type of cluster. This is an example of PNRCT structure because families within the treatment condition were nested within groups (one facilitator per group), while participants in the control condition were not.

Lohr et al. (2014) describe the PNRCT model as a hybrid model because the treatment arm includes a cluster structure, whereas the control arm remains unclustered. The analysis of such a data structure requires a combination of a two-level regression model for the treatment arm and a single regression model for the control arm.

I followed Lohr et al.'s (2014, p 48-49) model description to present the PNRCT model. However, I slightly modified this notation in order to follow Raudenbush and Bryk's (2002) notation. First, I used *i* for students and *j* for clusters. Second, I used r_{ij} and u_{1j} to denote the student random effect and the cluster random effect in the treatment arm. In addition, I used τ_{11} and σ^2 rather than σ^2_θ and σ^2_T (or σ^2_C) for the cluster variance and individual variance.

To illustrate this notation, I begin with the control arm, in which students are not assigned to a cluster. Let Y_{i0} denotes the student test score of student i in the control arm, with "0" indicating that students do not belong to a cluster. These students have no variability at the cluster level, which implies no u_{ij} effect. The subscript i refers to students, with $i = 1$ to n_C for the control arm where n_C is the control arm sample size. The model for the control arm can then be modeled in a single regression:

$$Y_{i0} = \beta_0 + r_{i0}, \tag{1.1}$$

in equation (1.1), β_0 is the mean score for students in the control arm and r_{i0} is the unique student random effect ($r_{i0} \sim N[0, \sigma^2]$). Notice that this model does not include a covariate indicating the treatment condition because all students belong to the control arm (e.g., $T_{i0} = 0$).

On the other hand, in the treatment arm students are assigned to a cluster. Here, Y_{ij} denotes students' outcomes such as test scores; thus, Y_{ij} implies student test score i from cluster j, for $j = 1$ to J. This is the same notation used in a two-level model of HLM (Raudenbush & Bryk, 2002). The subscript i refers to students, with $i = 1$ to n_j for the treatment arm students in cluster j. Then, the total number of students in the treatment arm is $n_T = \sum_{j=1}^{J} n_j$. The model in the treatment arm is $Y_{ij} = \beta_{0j} + \beta_{1j} + r_{ij}$. The variability of β_{0j} is not important in this model due to the fact that treatment effect (β_{1j}) is what matters. Thus, β_{0j} is represented as β_0. In addition, β_0 in the treatment arm is the same as β_0 in the control arm because random assignment in the PNRCT structure is expected to create two arms (treatment and control) with the same characteristics. Thus, the mean of the control arm should be the same as the mean of the treatment arm. Therefore, the model for the treatment arm is

Level 1:

$$Y_{ij} = \beta_0 + \beta_{1j} + r_{ij}. \quad (1.2)$$

Note that this model also does not have a covariate indicating the treatment condition because all students in this model are in the treatment arm (e.g., $T_{ij} = 1$).

Level 2:

$$\beta_0 = \gamma_{00}, \quad (1.3)$$

$$\beta_{1j} = \gamma_{10} + u_{1j}. \quad (1.4)$$

The combined model is

$$Y_{ij} = \gamma_{00} + \gamma_{10} + u_{1j} + r_{ij}. \quad (1.5)$$

In this model $\gamma_{00} + \gamma_{10}$ is the mean score for students in the treatment arm, r_{ij} is the unique student random effect ($r_{ij} \sim N[0, \sigma^2]$), and u_{1j} is the unique cluster random effect ($u_{1j} \sim N[0, \tau_{11}]$).

A hierarchical two-level linear model can be easily parameterized to model the PNRCT data structure (Talley, 2013). Models in equation (1.1) and (1.5) can be collapsed into a unified model by including an indicator variable for students' treatment status (Lohr et al., 2014). Let T_{ij} be the indicator variable, which takes the value of 1 ($T_{ij} = 1$) if students appear in cluster j of the treatment arm, and 0 if students appear in the control arm. Note that the treatment arm contains J clusters, $T_{ij} = 1$ for all students for $j = 1$ to J, and $T_{i0} = 0$ for j = 0. Then, it follows that j = 0, 1,…,J. The hierarchical two-level model is

Level 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}T_{ij} + r_{ij}. \quad (1.6)$$

Level 2:

$$\beta_{0j} = \gamma_{00}, \quad (1.7)$$

$$\beta_{1j} = \gamma_{10} + u_{1j}.$$

Note that the last equation (without numbering) is the same equation as the simple level two model above (1.4). By collapsing equations (1.6), (1.7) and (1.4) the combined model is obtained:

$$Y_{ij} = \gamma_{00} + \gamma_{10}T_{ij} + u_{1j}T_{ij} + r_{ij}. \quad (1.8)$$

In equation (1.8) the cluster effect appears only when T_{ij} =1 for j=1 to J and does not appear when T_{ij} = 0 for j = 0. This reflects the partially nested structure of the PNRCT model. A more flexible model allows σ^2 to be different for the treatment and control arms in which case $r_{ij} \sim N(0, \sigma^2_T)$ for students in the treatment arm (j = 1 to J) and $r_{i0} \sim N(0, \sigma^2_C)$ for students in the control arm.

The aforementioned PNRCT model can be called a "pure" PNRCT because it assumes only one treatment group with several clusters within it and only one control group with no clusters. However, the PNRCT model can be extended to other situations that will be discussed in Chapter 2 of the present study.

1.2.PNRCT Model Assumptions

Standard regression models are subject to a set of assumptions: error independence, homocesdasticity of variance, and normal error distribution. In standard regression models, when these assumptions are moderately or severely violated, parameter estimates are not automatically impacted.

However, the quality of parameter estimates and their standard errors may be affected. Thus, statistical techniques are used to minimize the impact (Fox, 2008). Like any other regression model, the PNRCT model is subject to a set of assumptions. Since the PNRCT model is a hybrid of an IRCT and a CRCT (Lohr et al., 2014), the CRCT arm has the structure of HLM. Thus, the PNRCT model is subject to the same model assumptions as HLM. These assumptions can be summarized as follows: first, the level 1 residuals are independently and normally distributed with mean zero and variance equal to σ^2 ($r_{ij} \sim N[0, \sigma^2]$). For the PNRCT model, this assumption applies when the treatment and control are assumed to have the same residual variance. However, some researchers may assume that treatment and control arms, of the PNCRT model, do not share the same residual variance. These residuals still have a mean equal to zero but the variances have different values. The previous section explained this. Second, level 1 and level 2 residuals are independent ($Cov[r_{ij},u_{1j}] = 0$). Third, random errors are independent among level 2 units ($u_j = [u_{0j}, \ldots, u_{Qj}] \sim iid\ N[0,\mathbf{T}]$). Note from equation (1.6) to (1.8) that at level 2 only one error term exists. Thus, the multivariate normal distribution, that applies to HLM, is reduced to a univariate normal distribution assumption with a mean of zero and a variance of τ_{11}. In other words, $u_{1j} \sim N[0, \tau_{11}]$). In addition, the u_{1j} does not need to be independent from the other level 2 errors because it is the only level 2 error in the model.

As in HLM, additional assumptions may apply to the PNRCT model when this is adjusted by covariates at level 1 and level 2. The first additional assumption is that level 1 predictors and r_{ij} are uncorrelated ($Cov[X_{qij}, r_{ij}] = 0$).

The second additional assumption is that level-2 errors are multivariate normally distributed, each with a mean of zero and a variance τ_{qq}. This assumption will apply only if there are at least two level 2 equations with error terms. In addition, they have a covariance $\tau_{qq'}$. Third additional assumption, the level 2 predictors are not correlated with level 2 error terms (Cov(W_{sj}, u_{qj})=0). Finally, predictors at each level are uncorrelated with the random effects at other levels (Cov [X_{ij},u_{qj}] = 0 for all qq' and Cov[W_{sj},r_{ij}] = 0) (Raudenbush & Bryk 2002).

1.3. Research Problem

The interest and focus of the present research is on how symmetric heavy-tailed distributions, such as t distributions affect a PNRCT model. More specifically, I am interested in how the violation of normal error distribution at level 2 negatively impacts the quality of the parameter estimates.

Roberts and Roberts (2002) and Lee and Thompson (2005) were the first researchers to advocate for the PNRCT approach by adapting its nonparallel structure. In 2008, Bauer et al. presented a model that more suitably met the requirements of the PNRCT design. However, since then few studies have used PNRCT models (Sander, 2011).

A small number of related Monte Carlo studies have been conducted to evaluate the adequacy of the PNRCT model. This model has been evaluated when competing with the ANOVA model (Baldwin, 2011; Korendijk, Maas, Hox, & Moerbeek, 2012), the fixed-effect approach model, and with a fully nested model assuming equal variances between conditions (Baldwin, 2011).

In addition, the model has been evaluated against a fully nested model specifying control subjects as clusters of size one (Korendijk, Maas, Hox, & Moerbeek, 2012; Sanders, 2011), a fully nested model specifying control subjects as one big cluster (Korendijk, Maas, Hox, & Moerbeek, 2012; Sander, 2011), and with a model with pseudo-clusters in the control condition (Sander, 2011). Some extensions of the PNRCT models have also been evaluated: a partially cross-classified model (Luo, Cappaert and Ning, 2015), a PNRCT three-level model with a higher hierarchy, and a PNRCT three-level model with a lower hierarchy (Tesller, 2014) have been evaluated competing with the full cross-classified model, a full three-level model and a full three-level model with observations nested within individuals, respectively. All these studies have included several conditions such as cluster size, number of individuals per cluster, ratio of imbalance, ICC values, methods of estimation and other factors for assessing the sensitivity of parameter estimates and associated significant tests, as well as Type I error rates, power rates and other related statistics. The relatively few studies on PNRCT so far have been performed assuming that model assumptions hold, which may not be true when working with real data.

When fitting linear regression models, it is recommended one assess the model assumptions (Fox, 1991; Raudenbush & Bryk, 2002). One of these assumptions is the normality of error distribution. This assumption is particularly important, because when working with real world data, normality distribution is not as common as one would think (Micceri, 1989). Using real data, and particularly educational data, the residual distributions can deviate from normality (Harwell & Gatti, 2001; Mass and Hox 2004a),

especially if the sample size is small, which may affect the quality of parameter estimates (e.g., standard errors) (Ketelsen, 2014).

Researchers sometimes decide to ignore departures of error distribution from normality, appealing to large sample theory (i.e. central limit theorem) (Fox, 1991; Lee, Nelder, & Pawitan, 2006) or trusting the robustness of specific methods (Pearson, 1931, Verbeke & Lasffre, 1997). Fox (1991) discussed three reasons why researchers must be concerned about non-normal error distribution in standard regression models. First, the efficiency of estimators is reduced. This occurs because least-square estimation is not robust in efficiency when non-normality is present. Additionally, the efficiency of estimators decreases substantially when the error term distribution is heavy-tailed. Second, a highly skewed error distribution compromises the interpretation of the least-squares fit. The interpretation is based on a conditional mean, which under highly skewed distribution does not accurately measure the center of the distribution. Finally, multimodal error distribution suggests the omission of at least one variable, implying that the regression model is misspecified.

The violation of the normality assumption in nested data is frequently present in applied settings (Delpish, 2006; DiPret, et al 1994), including data in HLMs and PNRCT models (Ketelsen, 2014). Thus, checking this assumption is important because in these models (HLMs, and PNRCT) the assumption of normal error distribution underlies both level 1 and level 2 (Bauer, et al. 2008; Mass & Hox, 2004a), and since the level 2 sample size, by definition, is smaller than the level 1 sample size, determining the effects of non-normal error distribution is necessary for hypothesis testing to produce accurate results.

Despite the fact that a large number of researchers have explored the effect of violating the assumptions of normal residual distributions in standard regression linear models (e.g., Glass, Peckham, and Sanders, 1972; Maravina, 2012; Peña, Zamar, & Yan, 2008), much less attention has been paid to HLM (Delpish, 2006; Shieh, 1999; Shieh, Fouladi, & Pullum, 2001) and other nested models such as the PNRCT model. Just recently, researchers have gone in depth on this issue with HLMs (Ketelsen, 2014).

According to Mass and Hox (2004b), some simulation studies have indicated that the violation of the normality assumption has almost no impact on regression coefficients and their standard errors, but variance components and their standard errors could be extremely biased. Raudenbush and Bryk (2002) asserted that failing to achieve the normality assumption at level 1 does not bias the estimation of the level 2 effects, but non-normality at level 2 will bias the standard errors at both levels. In regards to level 2, Raudenbush and Bryk stated that when fitting HLMs with educational data, heavy-tailed error distributions are usually produced. When this situation happens, the fixed effects will not be biased, but hypothesis testing and confidence intervals based on normality may be compromised, especially in the presence of outliers. Although Raudenbush and Bryk more specifically indicated the level at which non-normality is present, they did not provide any further information regarding the type of heavy-tailed distributions. However, they did cite Seltzer (1993), who used *t*-distributions with four and eleven degrees of freedom in his simulation study.

PNRCT models have received no attention regarding the violation of the normality of error distribution, and there is no clear explanation as to why this topic has not been investigated. Readers may infer that this is because the PNRCT models have been

recently proposed, or because these models are related in some way to HLMs, and thus the PNRCT models may be subject to the same negative implications that HLMs face when violating the normality assumption. However, research studies may present contrasting findings on the same topic. For that reason, we cannot assume that PNRCT models are affected in the same way as HLMs when the normality assumption is violated.

The question of how the PNRCT model's outcomes (e.g., fixed effects) are impacted when the normality assumptions at level 2 are violated is especially important. This is because remedial measures such as trimming clusters, performing a nonlinear transformation (on all model variables), or adding predictors at level 2 to account for non-normality are not common practices for correcting the violation of normality at level 2, and may not accurately account for the non-normal level 2 residual distribution. In addition, some asymptotic tests which rely on the normality assumption are applied at level 2, and non-normality may impact these estimates.

The interest and focus of the present research is on how symmetric heavy-tailed distributions, such as t distributions, affect a PNRCT model (fixed effects, Type I error rate and Power)[1]. This topic is particularly important when fitting nested models for some reasons. First, in educational data, heavy-tailed distributions are more common than one would expect. Micceri (1989), for instance, found that among 440 large-sample achievement and psychometric measures, 49.1% had at least one heavy-tailed distribution. Second, the efficiency of least-square parameter estimates decreases and gives rise to outliers (Fox, 1991). Third, several statistical packages, such as R (Bates,

[1] This research does not focus on random effects (U_{1j}) for two reasons. First, the PNRCT model may use fewer clusters than traditional HLM. Second, it is well documented that random effects are estimated by using ML, which relies on the assumption of normal error distribution (Meijer, van der Leeden, & Busing, 1995). Therefore, when normal assumptions do not hold, the accuracy of the variance components estimates is already compromised.

Mächler, Bolker, & Walker, 2014), use ML estimation for estimating fixed effects in HLM, and ML estimation assumes a normal distribution, so when normality is not achieved, complications may arise (Delpish, 2006). Fourth, no study evaluating the impact of a heavy-tailed error distribution at level 2 on a PNRCT model's estimates and statistical tests (fixed effect, Type I error rate and Power) has been performed.

Chapter 2

Literature Review

2.1. Randomized Controlled Trials and Method of Analysis

Randomized Controlled Trials (RCT) are not a common practice in education although articles using RCT can be found in research journals. Two important factors potentially limiting the use of RCT are policymakers' and researchers' perceptions that such studies are impractical due to cost and administrative complexity, and the ethical issue of denying an intervention to any subject (Coalition for Evidence-Based Policy, 2007; Orr, 1999). However, RCT permits researchers to provide causal estimates when determining the impact of educational interventions, thus making it fundamental to the practice of best evidence in education. RCT provides an equal and independent chance for individuals to be assigned to a treatment or control condition, minimizing or eliminating the effects of confounding variables, which create conditions for determining treatment effects. In an optimal scenario, the only difference between treatment and control groups in RCT should be exposure to the intervention.

The process of RCT can be carried out in different ways, for instance, from traditional methods such as drawing the names of individuals from a box or flipping a coin (Gertler et al., 2011) to technological methods such as computer programs (Converdale et al., 2013). Regardless of the methods used, researchers need to guarantee that subjects will be randomly assigned to treatment and control groups in a rigorous manner.

Although randomly assigning subjects to control trials seems to be an easy process, there are several issues that compromise this approach. This idea is supported by Coverdale et al. (2013), who argued that RCT's cost time and money and are challenging in practice because of complexities associated with these designs as well as ethical constraints. Moreover, non-compliance to the assigned intervention also complicates RCTs (Caliendo, 2006) and occurs when individuals abandon the treatment condition (for any reason) or subjects from the control group exhibit preferences for or signs of receiving the treatment (Roberts, Geppert, Connor, Nguyen, & Warner, 2001; Roberts, Geppert, Coverdale, Louie, & Edenharder, 2005).

Despite the difficulties with RCT, it has several advantages when it is accurately implemented. Burtless (2002) stated four main advantages of RCT: RCT (a) creates the conditions for identifying the effects of treatment with precise reliability; (b) it eliminates systematic relationships between treatment status and observed and unobserved characteristics of participants; (c) it permits researchers to measure the effects of environmental changes not previously observed; and (d) it makes results convincing and understandable to other researchers and policymakers. Furthermore, as Boruch et al. (2002) notes, RCT permits researchers to make comparisons between two groups with similar characteristics and make valid statements regarding the results. Three approaches of RCT are (a) Individual Randomized Controlled Trials (IRCT), (b) Cluster Randomized Controlled Trials (CRCT), and (c) Partially Nested Randomized Controlled Trials (PNRCT) (Bauer et al., 2008; Lohr et al., 2014).

Under IRCT, individuals (e.g., students) are randomly assigned to either control or treatment groups (Gertler et al., 2011; Lohr et al., 2014). To illustrate the nature of the IRCT design, consider the following example. Zepeda, Richey, Ronevich, and Nokes-Malach (2015) studied the impact of metacognitive skills on students' motivation, learning and future learning in classrooms. These researchers randomly assigned 46 eighth-grade students into treatment or control conditions. In the treatment condition, students (n_1 = 23) received extensive problem-solving practice. Meanwhile, in the control condition, students (n_2 = 23) received more limited problem-solving practice along with metacognitive instruction and training.

When IRCT is used, as in the previous example, the data can be analyzed by using a simple linear regression. The treatment effect then is estimated by the slope associated with the treatment variable and tested against zero with a t-test (Caliendo, 2006; Cerulli, 2015). When a linear regression is used researchers may include one or more covariates in the model to increase the precision of the estimates (Cerulli, 2015), which may increase statistical power (Gail, Wieand, & Piantadosi, 1984; Raab & Butcher, 2001; Robinson & Jewell, 1991).

Equation (2.1) and equation (2.2) are the mathematical representations of a simple linear regression and multiple linear regression, respectively:

$$Y_i = \beta_0 + \beta_1 T_i + e_i \tag{2.1}$$

and

$$Y_i = \beta_0 + \beta_1 T_i + \sum_q \beta_q X_{iq} + e_i \,, \tag{2.2}$$

where Y_i is the outcome variable, T_i is the treatment condition for the *ith* student, β_0 is the intercept of the model, β_1 is the treatment effect, and β_q captures the effect of the X_{iq} covariate. Finally, e_i is a unique effect associated with the *ith* student.

Contrary to IRCT, CRTC does not randomly assign students. Rather, CRTC randomly assigns a higher level of hierarchy (e.g., school, districts) into treatment or control conditions, and every individual (e.g., students) within the cluster (e.g. school, district) receives the same condition (Raudembush & Bryk, 2002). An example of CRTC design is the research performed by Lesaux, Kieffer, Kelley and Harris (2014). In this study, the language and literacy skills of linguistically diverse sixth grade students (N = 1469), taught by 50 teachers, were examined. The researchers randomly assigned the 50 teachers, and consequently their students, into one of two conditions. Twenty-five teachers were assigned to the treatment condition and the other 25 were assigned to the control condition. The treatment condition was a 20-week program that attempted to improve students' vocabulary knowledge, morphological awareness skills, and comprehension of expository material. The control condition did not include any type of program. A common characteristic shared by all students in this study was that English was not a primary language at home.

When working with nested data structure, researchers have to be aware that observations within clusters are very likely to be correlated. If the nested structure is not taken into account, the assumption of independent observations would be violated. Fortunately, the nested structure of data has become relatively easy to handle because of the development of Hierarchical Linear Modeling (HLM) (Raudenbush & Bryk, 2002). The increasing popularity of Hierarchical Linear Models has also given rise to

educational researchers' use of CRCT. Not only does HLM have the capacity to control for the correlation created among observations within classrooms, but researchers can also model cross level interactions, partition the variances and covariances components, and improve estimation of individual effects (Raudenbush & Bryk, 2002). Therefore, researchers more frequently select classrooms or teachers for receiving a treatment rather than students. The following equations represent an HLM framework for CRCT data analysis:

level-1

$$Y_{ij} = \beta_{0j} + r_{ij} \quad (2.3)$$

and level-2

$$\beta_{0j} = \gamma_{00} + \gamma_{01} T_j + u_{0j}. \quad (2.4)$$

In these equations Y_{ij} is the outcome variable, β_{0j} reflects the *jth* classroom (or teacher) mean, γ_{00} represents the grand mean, γ_{01} captures the treatment effect, T_j is the treatment condition of the j_{th} classroom, and r_{ij} and u_{0j} are the unique effects associated with the *ith* student and the *jth* classroom (or teacher), respectively. In addition, r_{ij} and u_{0j} are assumed to be normally distributed with a mean of zero and variance of σ^2 and τ.

Researchers may include covariates at both levels of the model to improve precision of the estimates, explain sources of variability in treatment effects (Raudenbush & Bryk, 2002) and increase power (Gail, Wieand, & Piantadosi, 1984; Raab & Butcher, 2001; Robinson & Jewell, 1991). In such a case, the models to be used are represented in equations (2.5) and (2.6):

Level-1

$$Y_{ij} = \beta_{0j} + \sum_q \beta_{qj} X_{qij} + r_{ij} \quad (2.5)$$

and level-2

$$\beta_{0j} = \gamma_{00} + \gamma_{01} T_j + \sum_p \gamma_{pj} W_{pj} + u_{pj}. \quad (2.6)$$

β_{qj} is the qth regression coefficient (p = 0, 1, 2, …,q) for the *jth* classroom or teacher, and captures the impact of the *qth* student-level predictor X_{qij}, γ_{pj} is a slope capturing the impact of the classroom-level predictor W_{pj}, and r_{ij} is the level-1 residual and u_{pj} is the residual for the level-2 model (Raudenbush & Bryk, 2002). r_{ij} and u_{pj} are assumed to be normally distributed with a mean of zero and variance of σ^2 and τ, respectively. All other terms were previously defined.

A third type of RCT is PNRCT. The PNRCT design and the model (equation [1.6] to equation [1.8]) to analyze its data were explained in Chapter 1. As noted earlier, in PNRCT randomization is done on an individual basis, distributing individuals into treatment and control conditions. Then, individuals in the treatment group receive the intervention in clusters. They may or may not be randomly assigned to these clusters. The treatment is administered in a group setting so that multiple individuals receive the treatment together. On the other hand, in the control group students remain unclustered (Bauer et al., 2008; Lee & Thompson, 2005; Lohr et al, 2014). Examples and the equations to analyze the PNRCT data were provided in Chapter 1, Section 1.1.

2.2. When is the PNRCT Design Possible?

The PNRCT design is probably not the design of choice in most educational settings because of two conditions. The first condition is that in educational experiments the units of randomization are typically classrooms or clusters rather than students.

Thus students in the control condition are still clustered within classrooms, generating a cluster effect (Tesller, 2014). Additionally, the nature of the PNRCT design may produce a high likelihood of contamination in the control condition if an experiment is conducted in one school. This is because students in the treatment and control conditions could belong to the same school or the same classroom, depending on the type of experiment.

The PNRCT design may, however, be a good approach in other situations. For instance, this model is often more suitable in educational settings such as extracurricular programs, which can be classified as before-school, after-school, and summer learning programs. Lohr et al. (2014) declared that PNRCT design would typically be used when evaluating the effect of these types of programs. Notice that when using the PNRCT design in extracurricular programs the likelihood of contamination may decrease because treatment and control students do not necessarily belong to the same school or classroom.

The following example illustrates a hypothetical after-school program in which researchers evaluate whether or not a program helps low-achieving students to improve their reading and mathematics skills using the PNRCT design. The target population of the study is students in fourth to the sixth grade who need supplementary instruction to enhance their reading and mathematics skills. The researchers recruit students identified by teachers as needing supplemental academic support from different schools. Students who sign up for the study agree to finish the academic year. Students are then randomly assigned to treatment and control conditions. Students who are in the treatment condition are distributed into different instructional centers to receive the intervention while students in the control condition are put on a waiting list.

The intervention takes place over a considerable period of time. At the end of the intervention, treatment and control students take a reading and a mathematics test and researchers determine the treatment effect using the PNRCT model.

This hypothetical example represents a PNRCT design because (a) students were assigned to treatment and control groups, (b) within the treatment group students were assigned to clusters, and (c) students within the control group remained unclustered.

The PNRCT design may be used in other applied settings. For instance, take another hypothetical scenario in which pedagogical mathematics software is evaluated. The researchers invite students from different districts to participate in the study. After recruiting the participants, half of them are randomly assigned to use the software. These students are assigned to different computer centers in which an instructor trains them in the use of the software and answers any questions. The other half of the students are in the control group. These students receive different types of materials, such as books and handouts, with the same content as in the software. In addition, the students in the control group are told that they will study the material individually in their homes. At the end of the intervention, students are evaluated by using a mathematics test. This experiment is a PNRCT design because of the same three facts mentioned in the previous example. Thus the outcomes of such a design can be analyzed by using equation (1.8).

Although the PNRCT model was developed for the PNRCT design, which is an experimental design, the PNRCT model may be used in quasi-experimental designs. Lohr et al. (2014) argue that experimental and quasi-experimental designs involve the same issues, thus they assert that the PNRCT model can be used in quasi-experimental designs, especially in settings in which matching is used.

In addition, Lohr et al. believe that it is possible to use the PNRCT in regression discontinuity designs and with instrumental variables. However, Lohr et al. advise researchers to use a large sample size in order to attain the same level of statistical precision.

Consider the following hypothetical example for the use of the PNRCT model in a quasi-experimental design. A researcher received a data set containing data from one school. In this school, 200 students were separated into two groups (n_1 = 60 and n_2 = 140). No random procedure was used to separate students. To claim causal inferences, the researcher used a propensity score matching technique to select 60 students from n_2 who were supposed to have the same characteristic as n_1. The two groups worked in different classrooms. In both groups, students received the same science material. However, in one group (n_1) students were randomly assigned to subgroups with five students each and they worked in a cooperative learning context. Students in the other group worked individually. At the end of the week, researchers administered a science test to the students and a survey requiring demographic information.

2.3. Some Considerations about the PNRCT Model

There are some further considerations regarding "pure" PNRCT design that must be noted. First, there can only be one treatment. The fact that the treatment condition has clusters does not imply the existence of different treatments. The same treatment is given to all subjects in the treatment condition regardless of cluster assignment.

However, since the treatment may be delivered by different subjects at different levels of hierarchy (e.g., instructors deliver to participants, teachers deliver to students, therapists deliver to patients), a random effect might be introduced to account for cluster dependency in level 2 of the model. This random effect may be present "due to the particular composition of the group, the fidelity of implementation of the treatment protocol, the effectiveness of the treatment administrator for the group, or other factors" (Bauer et al., 2008, p. 8).

Therefore, the notation $T_{ij}=1$ in equation (1.8) implies that the treatment is delivered to subject *i* within the cluster *j* of the treatment arm. The β_{1j} parameter, also in equation (1.8), reflects the treatment effect in the cluster *j* within the treatment condition. This is because of the existence of a random effect (u_{1j}), which permits the treatment cluster mean to vary across clusters within the treatment condition.

Second, "if the [PNRCT] design includes few clusters, it is difficult to learn about differences among clusters." (Baldwin et al., 2011, p, 162). If this situation arises, it is important to know how to proceed. A typical solution may be changing the research design to another type of analysis, such as to an ANOVA model. The problem is that, to my best knowledge, no research study has been conducted to determine when it is convenient to change from a PNRCT model to the ANOVA model or any other type of model, so this question remains unanswered.

To avoid the aforementioned problem, researchers should know a priori the number of clusters to include in the PNRCT design. Lohr et al. (2002) discuss in depth the required number of clusters for estimating the average treatment effect in the PNRCT. They argue that a power analysis must guide the selection of the number of clusters.

Other researchers have a variety of recommendations for the number of clusters to use. However, the context of these recommendations is in the HLM field. For example, Browne and Draper (2000) noted that when using between six and twelve clusters with a Restricted Maximum Likelihood (RML) estimation, a reasonable variance will result. On the other end of the spectrum, Busing (1993) suggested that for an accurate estimation of the variance more than 100 clusters are needed.

Mass and Hox (2004b) suggested that, if researchers are interested in fixed effects, ten clusters are enough to have good estimates. If the interest is contextual effects, 30 clusters will be sufficient. If researchers are interested in accurate estimates of standard errors, 50 clusters will be needed. These conflicting recommendations make the number of clusters to use at level 2 in an HLM unclear. Thus, researchers may use as many clusters as possible. This could also apply to the PNRCT field.

2.4. Traditional Methods for Analyzing Partially Nested Data Structures

The importance of analyzing PNRCT is to properly account for the partially nested structure because by failing to do so, or by using inappropriate models, serious complications may arise. These complications will be presented through this section.

Unfortunately, the importance of properly analyzing PNRCT in quantitative studies has not been widely documented (Bauer et al., 2008; Lee & Thompson, 2005; Lohr et al., 2014; Roberts & Roberts, 2005). A few authors have provided didactic papers introducing researchers to the analysis of such designs (Bauer et al., 2008; Lohr et al., 2014), but more work needs to be done.

An important question is whether researchers have failed to analyze PNRCT structure data with the appropriate model. Researchers have provided little evidence in the past years in this regard. For instance, Sanders (2011) reviewed four peer-reviewed journals: (a) American Education Research Journal (AERJ; first published in 1964), (b) Contemporary Educational Psychology (CEP; first published in 1910), (c) Journal of Educational Psychology (JEP, first published in 1976), and (d) Remedial and Special Education (RASE, first published 1974).

This researcher found 75 research articles (16%) among 467 where researchers used randomized experiments. Of these 75 articles, ten articles (13%) presented PNRCT designs, and researchers correctly analyzed this data structure in only two articles.

To estimate the prevalence of PNRCT design in current educational research, I reviewed research articles published in three peer-reviewed educational research journals from 2011 to 2015. The journals reviewed were the AERJ, Evaluation Review (ERX; first published in 1977), and CEP. These journals were selected for review because they cover a wide range of topics in education and the social sciences, and they are known as prominent journals with well-established and wide readerships (Sanders, 2011).

Researchers used randomized controlled trials as part of their research in only 67 out of 437 articles (15%). The PNRCT design was used in eleven out of these 67 articles (16%). However, in none of them was the PNRCT model used for analyzing data. Precisely characterizing the frequency of studies that use PNRCT design may not be possible because the studies under analysis are only those that have been published and excludes unpublished studies such as dissertations and conference papers.

This evidence suggests that researchers are likely to employ traditional methods of analysis even when these methods are not optimal because of their unfamiliarity with partially nested data analysis techniques. This is likely because PNRCT models are a relatively new topic and PNRCT data analysis techniques have thus far received little methodological attention. Consequently, researchers have not been exposed to examples that correctly analyze PNRCT data.

This section presents several models that researchers have used to analyze PNRCT data structure. These models are separated into two main groups: models that do not take into account the PNRCT data structure and models that do. In most of the models, the original notation has been modified in order to be consistent. To be specific, I used the notation presented by Raudenbush and Bryk (2002), denoting individuals with the subscript *i* and denoting groups with the subscript *j*, rather than the notations found in the original articles. Readers should take this consideration into account if they compare the models of this study with models in the original articles.

2.5. Disregarding Clustered PNRCT Data Structure

Bauer et al. (2008) wrote that the most common approach used to analyze PNRCT is to ignore the clustered data structure in the treatment arm; this is consistent with Sanders (2011), who found that in 70% of research papers (four journals were reviewed between 2007 and 2009) with PNRCT data structure, researchers analyzed the data at an individual level, ignoring the clustered structure. Many statistical models can be used when ignoring clustered data structures, but the simple regression model (Baldwin et al., 2011; Bauer et al., 2008) and one-way ANOVA are probably those most commonly seen in the literature.

2.5.1. Simple regression models and one-way ANOVA. In education literature, researchers report simple regression models (SRM) as the most common method for analyzing PNRCT data (Bauer et al., 2008; Sanders, 2011). In almost every study involving PNRCT, researchers describe the problems of this model when analyzing PNRCT data, and we will continue with that tradition. The simple regression model is defined as follows:

$$Y_i = \beta_0 + \beta_1 T_i + e_i , \tag{2.7}$$

where T_i represents the treatment condition through an indicator variable that takes the value of one when subjects are in the treatment groups and zero for subjects that are in the control groups. When T_i in equation (2.7) takes the value of zero the model reduces to

$$Y_i = \beta_0 + e_i , \tag{2.8}$$

and when T_i takes the value of one, the model becomes

$$Y_i = \beta_0 + \beta_1 + e_i. \tag{2.9}$$

In these three equations, the parameter β_0 captures the group mean response for the control arm, and β_1 captures the treatment effect between the treatment and control arm. e_i captures the *ith* random error, and it is assumed to be normally distributed with a mean of zero and variance equal to σ_e^2 ($e_i \sim N[0, \sigma_e^2]$).

Another model that has been used for analyzing PNRCT data is one-way ANOVA (Sanders, 2011). As in other statistical models, this model describes the relationship between the outcome variable and the treatment conditions. The following mathematical form shows this relationship:

$$Y_{ij} = \mu + \tau_j + \varepsilon_{ij}. \quad (2.10)$$

In equation (4) Y_{ij} represents the *ith* observation on the *jth* group (treatment condition), μ is the overall mean, τ_j captures the treatment effect and ε_{ij} represents the random error for the *ith* observation in the *jth* group. In a one-way ANOVA, it is assumed that $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ and $\sum \tau_i = 0$.

All researchers who are interested in PNRCT agree that models that ignore clusters in the treatment arm present several problems (Bauer et al., 2008; Lee & Thompson, 2005; Lohr et al., 2014; Roberts & Roberts, 2005). The first implication, and the most obvious, is the violation of the independence assumption, especially in the treatment arm (Bauer et al., 2008). A second implication is that the model assumes an intra-class correlation (ICC) of zero, which may not be accurate since PNRCT design presents a nested data structure. By assuming an ICC of zero, when in fact it exists, the standard error for the fixed effects are erroneously estimated, which causes an inflation of the Type I error rate for the test of the intervention effect (Mass & Hox, 2004a).

Another implication is that these models only estimate the between-subject variance which is assumed to be the same for both the treatment and control conditions ($V(Y_i|T = 1) = \sigma_e^2$ for the treatment condition and $V(Y_i|T = 0) = \sigma_e^2$ for the control condition) although it is very unlikely that treatment and control conditions would have the same variance (Bauer et al., 2008). The between-cluster variance in the treatment arm then is ignored. This is pooled into a single term representing the between-subject variability. Therefore, this variance structure is inconsistent with the variance structure of PNRCT design.

2.6. Including Clustered Structure

Researchers have proposed several models that take into account the nested structure of the data of the PNRCT design (Baldwin et al., 2011; Bauer et al., 2008: Sanders, 2011). All of these models, except for Bauer et al., present several negative implications. This section attempts to provide a summary of the following models: a nested model with clusters as fixed effects, a fully nested model, a pseudo-clustering control condition model, and a model that treats the control arm as one large cluster.

2.6.1. Nested model with clusters as fixed effects. One approach for modeling PNRCT data is treating the cluster effects in the treatment arm as fixed (Baldwin et al., 2011). This model requires the inclusion of a set of indicator variables for the clusters within the treatment arm. The inclusion of such variables captures the cluster mean differences. An additional indicator variable is also needed for capturing the mean of the control condition. Mathematically, this model can be illustrated as follows:

$$Y_i = \beta_0 Control_i + \beta_1 Cluster1_i + \beta_2 Cluster2_i + \beta_3 Cluster3_i + \ ... + \beta_g Clusterg_i + e_i. \quad (2.11)$$

Control$_i$ is an indicator variable for the subjects within the control arm, *Cluster1* to *Clusterg* represent indicator variables (e.g., 1= cluster one, 0 = otherwise) and indicate that students are distributed among clusters in a particular way. Each β_g captures the effect associated with each cluster within the treatment arm. In this model, the overall intercept is not estimated. Instead, β_0 is estimated representing the mean of the control arm. Since the overall intercept is not estimated and there is no reference group, the model can estimate each β_g, capturing the mean for each cluster.

This model does not evaluate the treatment effect directly. Instead, the treatment effect is tested by the use of a contrast that combines the mean of the clusters within the treatment arm and the mean of the control group (e.g., the coefficient of the contrast would be -1 for the control group and 1divided by the number of clusters for clusters 1 to *g*). Additionally, this model assumes that the variance between subjects (σ_e^2) is a common variance for the treatment and control arms. Thus the variance of the treatment arm is equal to the variance of the control arm ($\sigma_{e0}^2 = \sigma_{e1}^2$), so $V(Y_i|Control) = \sigma_0^2 = \sigma_e^2$ and$V(Y_i|Treatment) = \sigma_1^2 = \sigma_e^2$.

The variance of the clusters in the treatment arm is defined as the summation of the variance between clusters plus the variance between subjects, which is $V(Y_i|Treatment) = \sum_{g=1}^{G} p_g \left(\beta_g - \bar{\beta}\right)^2 + \sigma_e^2$. Here, *G* is the total number of clusters in the treatment arm, p_j is the proportion of subjects for the corresponding cluster, and $\bar{\beta}$ represents the cluster grand mean computed as $\bar{\beta} = \sum_{g=1}^{G} p_g \beta_g$.

Although this model correctly depicts the variance structure of the PNRCT structure when the between-subject variance is the same for both the control and the treatment arms (Baldwin et al., 2011), the model is problematic if taking into account differences between clusters through fixed effects. First, cluster-to-cluster differences contribute to explained variance in the model, whereas the source of these differences may be unknown, assuming a random sampling of clusters. The Type I error rate for the test of the intervention effect is therefore only accurate when inferences are restricted to the specific clusters in the study (e.g., treatment Groups 1–4) (Serlin, Wampold, & Levin, 2003; Siemer & Joorman, 2003a, 2003b).

In contrast, one generally seeks to make inferences to the broader population of clusters (e.g., all possible treatment groups, not simply those in the study). For such inferences, the mean square error (*MSE*) of the model, $\widehat{\sigma}_e^2$ fails to fully represent the unexplained variance (is negatively biased), because cluster-to-cluster variance has been excluded. As when ignoring clustering, the consequence is that the test of the intervention effect will have a higher than desired Type I error rate (Baldwin et al., 2011, p.156).

2.6.2. Treating each individual in the control arm as a single cluster (fully nested model). Some researchers have stated that to deal with the partial structure of the PNRCT design, each individual in the control arm can be treated as a single cluster, which implies that each cluster in the control arm has only one individual (Baldwin et al., 2011; Bauer et al., 2008; Sander, 2011). Meanwhile, individuals in the treatment arm are grouped in clusters with more than one individual. To model this data structure, researchers have used a hierarchical linear modeling (fully nested model) approach by using equations (2.12) to (2.15).

Level 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}T_{ij} + r_{ij}, \tag{2.12}$$

Y_{ij} represents the outcome variable (e.g., posttest student scores) of subject *i* in cluster *j*. T_{ij} is the treatment condition. Similarly, β_{0j} represents the random cluster mean, and β_{1j} captures the fixed treatment effect for cluster *j*. Finally, r_{ij} is still the level-1 residual.

Level 2 of this fully nested model is

$$\beta_{0j} = \gamma_{00} + u_{0j}, \tag{2.13}$$

$$\beta_{1j} = \gamma_{10}. \tag{2.14}$$

Oftentimes, (2.12), (2.13) and (2.14) are combined to obtain a mixed model representation:

$$Y_{ij} = \gamma_{00} + \gamma_{10} T_{ij} + u_{0j} + r_{ij}, \quad (2.15)$$

where γ_{00} is the mean for the subjects within the control arm, u_{0j} is the random effect for the mean in the control arm, and γ_{10} captures the fixed treatment effect. Both r_{ij} and u_{0j} are assumed to be normally distributed with a mean of zero and variances σ^2 and τ_{00} (e.i. $r_{ij} \sim N(0, \sigma^2)$ and $u_{0j} \sim N(0, \tau_{00})$). Bauer et al. (2008) and Baldwin et al. (2011) stated that the variance structure for this model is the same for both the treatment and the control arms, which is $V(Y_{ij}|T = 1) = V(Y_{ij}|T = 0) = \tau_{00} + \sigma^2$, and the intra-class correlation (ICC, ρ) for both the treatment and control arms is assumed to be

$$\rho = \frac{\tau_{00}}{\tau_{00} + \sigma^2} \ . \quad (2.16)$$

Although fully nested models take into account the nested data structure, the model in equation (2.15) presents some complications. First, this model is not adequate for the PNRCT data structure, because within the control arm of the PNRCT, participants are not clustered in any type of structure. Second, the variance is divided into two components, between and within clusters, in both the treatment and the control arms. However, this partition is meaningless for the control arm, because each cluster in the control arm only has one individual; thus variability within clusters cannot exist. Third, since the variance is artificially divided between and within clusters, in the control condition, an ICC value of zero is not plausible even though each cluster only has one individual (Baldwin et al., 2011). Fourth, the cluster treatment effect at level-2 is not random, implying that all clusters in the treatment arm and all "clusters" in the control arm have the same effect, which may not be true.

Additionally, chances are that this model does not produce equal variances across conditions between subjects. If this is the case, this model will not accurately estimate the standard errors for testing the fixed effects.

The last problem could be fixed by allowing heteroscedasticity between treatment and control arms. If so, the variance structure is modified as $V(Y_{ij}|T = 1) = \tau_{00} + \sigma_1^2$ and $(Y_{ij}|T = 0) = \tau_{00} + \sigma_0^2$, where σ_1^2 is the variance in the treatment arm and σ_0^2 is the variance in the control arm. Baldwin et al. (2011) confirmed that this modification "conforms completely to the underlying variance structure of the data" (p. 152). Nevertheless, it still includes the meaningless partition of the variance, although the components of the variance are easy to interpret.

2.6.3. Pseudo-clustering control condition. One strategy for handling PNRCT data structures is randomly assigning observations to create pseudo-clusters in the control arm (Sanders, 2011). Thus, the data analysis can be performed by using HLMs (fully nested models). The mathematical representation is the same as in equations (2.12) to (2.15). The assumption of the residuals at levels one and two are still $r_{ij} \sim N(0, \sigma^2)$ and $u_{0j} \sim N(0, \tau_{00})$. In this model the ICC is meaningful for both control and treatment arms.

This model might present different variances between the treatment and the control arms. When this is the case, the assumption of homogeneity of variance can be relaxed (Sanders, 2011). Then the structure of the residual variance changes to

$$\ln(\sigma^2) = \alpha_0 + \sum \alpha_1 C_j, \qquad (2.17)$$

where C_j is a predictor from level-1 (Raudenbush & Bryk, 2002). This implies that the variance heterogeneity can be modeled.

Although artificially creating clusters in the control arm of the PNRCT data structure seems to be a good solution for handling this type of data structure, this solution is not free of negative implications for the model. The problems occur when subject observations within the pseudo-clusters are uncorrelated. First, the power for the treatment effect will decrease because this test depends on degrees of freedom, which comes from the number of clusters and not from the individuals. Second, the Type I error rate might be distorted because residuals in the control arm are likely to be larger than the treatment arm once clusters are taken into account (Sander, 2011). Another important drawback is that different researchers could create different clusters for the same data leading to different statistical results.

2.6.4. Treating the control arm as one large cluster. Sander (2011) evaluated the possibility of treating the control arm as one large cluster. The model specifications and negative implications are the same as the previously discussed for the pseudo-clustering control condition model.

2.7. Partially Nested Model

As previously mentioned, the PNRCT model is a hybrid model because it is a combination of a simple regression model and a hierarchical regression model. The PNRCT differs from a two-level HLM by not allowing random effects for the intercepts. At the same time, the treatment slope condition allows a random effect to be estimated and tested. Bauer et al. (2008) presented this model that meets the requirements of the PNRCT design, provides a valuable test of the whole treatment effect, and permits the proper determination of the degree to which the treatment effects vary over treatment

clusters within the treatment arm. The model was mathematically represented by equation (1.8). However, the equation is presented here one more time:

$$Y_{ij} = \gamma_{00} + \gamma_{10}T_{ij} + u_{1j}T_{ij} + r_{ij},$$

T_{ij} is an indicator variable, which takes the value of 1 ($T_{ij} = 1$) if students appear in cluster j of the treatment arm, and 0 if students appear in the control arm. Note that the treatment arm contains J clusters, $T_{ij} = 1$ for all students for $j = 1$ to J, and $T_{i0} = 0$ for j = 0. Then, it follows that for $j = 0, 1,\ldots,J$. In this equation, the cluster effect appears only when $T_{ij} = 1$ for $j=1$ to J and does not appear when $T_{ij} = 0$ for $j = 0$. This reflects the partially nested structure of the PNRCT model.

In this model, when T_{ij} takes the value of zero, a model for the control condition is obtained. This is mathematically represented by equation (2.18):

$$Y_{ij}|(T = 0) = \gamma_{00} + r_{ij}, \qquad (2.18)$$

and when T takes the value of one (treatment condition) , the equation is

$$Y_{ij}|(T = 1) = \gamma_{00} + \gamma_{10} + u_{1j} + r_{ij}. \qquad (2.19)$$

Notice that r_{ij} and u_{1j} are assumed to be independent and normally distributed: $r_{ij} \sim N(0, \sigma^2)$ and $u_{1j} \sim N(0, \tau_{11})$. Here, σ^2 represents the variance at level one and τ_{11} the cluster variance in the treatment arm. Note that the model in equation (2.18) is a simple regression model just for the control arm, and the model in equation (2.19) is an HLM just for the treatment arm.

One important difference between HLMs and the PNRCT model is the ICC. In PNRCT, the ICC only exists for the treatment arm, but not for the control arm. This happens because in the control arm clusters do not exist. Thus, the ICC for the treatment arm is

$$ICC_{treatment} = \frac{\tau_{11}}{\tau_{11}+\sigma^2}. \quad (2.20)$$

Additionally, in the PNRCT model the ratio of the variance between the treatment arm and the control arm is defined as

$$\frac{\sigma^2}{\tau_{11}+\sigma^2} = 1 - ICC_{treatment}. \quad (2.21)$$

τ_{11} represents the variance of the treatment slopes, and each slope is the difference between the conditional means (i.e. the treatment mean minus control mean).

Two of the most important features of the PNRCT model are that it better fits the data structure of the PNRCT design, and it provides estimates with interpretations that are more accurate than those of previous models. This model also takes into account the correlation between observations within clusters in the treatment arm. The model is able to capture variance heteroscedasticity (if required) between the control and treatment arms and tests whether treatment membership is a significant predictor of the outcome (Bauer et al., 2008). These facts make Bauer et al.'s (2008) approach the most accurate model for properly handling data from PNRCT.

2.7.1. Unconditional PNRCT model. The first step in the analysis of HLMs is usually fitting an unconditional (no covariates) model to estimate the variance components associated with each factor and to test them against zero (Raudenbush & Bryk, 2002). A statistically significant variance component signals that there is variation to be explained, whereas a non-significant component usually means that the factor is treated as a fixed effect or is dropped altogether from the model. Researchers may use an unconditional model for the treatment arm because they expect correlated observations after the treatment has been administered. This is because subjects are supposed to receive the treatment within the clusters in the treatment arm. However, this action

represents an additional step. Equation (2.19) represents the unconditional model, since it takes into account only the subjects that are within clusters in the treatment arm. As mentioned earlier, this model produces σ^2 and τ_{11} allowing us to estimate the ICC as in equation (2.20).

2.7.2. PNRCT model with heterogeneous variance. The PNRCT model can be fitted with homogeneous or heterogeneous variance across arms. Since the variance of r_{ij} between the treatment and the control in equation (1.8) may not be the same for the treatment and control arms, researchers may fit a model with heterogeneous variance (Lohr et al., 2014; Roberts & Roberts, 2005). For example, a model with heterogeneous variance can be specified by modifying the assumption in $r_{ij} \sim N(0, \sigma^2)$, which will produce a model with heterocedasticity between arms (Bauer et al., 2008). Thus the model will produce

$$r_{ij}|(T = 0) \sim N(0, \sigma^2_{Control}) \tag{2.22}$$

and

$$r_{ij}|(T = 1) \sim N(0, \sigma^2_{Treatment}). \tag{2.23}$$

If the model has homogeneous variance at level-1, the variance is

$$Var(r_{ij}) = \sigma^2. \tag{2.24}$$

2.7.3. PNRCT model with covariates. Random assignment usually guarantees a strong counterfactual. However, this is not always possible, because in practice internal consistency is threatened by many factors. Thus, including covariates in the analysis may represent a more realistic situation.

These variables are included in the model when analyzing the data for two reasons: (a) for improving the precision of the estimates and statistical power, and (b) for adjusting residual treatment-control baseline characteristics (Lohr et al., 2014). Additionally, in many cases, researchers might be interested in controlling for preexisting differences between individuals (Bauer et al., 2008), or may be focused on how the extent subject achievement-change in the treatment group differs from the control group. The first interest is satisfied by including pretest scores as a covariate in the model. The second interest may be satisfied by using a gain score variable (pretest score subtracted from posttest score) (Lohr et al., 2014).

PNRCT models are able to handle explanatory variables either at level 1 (e.g., students) or level 2 (e.g., teachers, schools, or any other higher level of hierarchy) (Bauer et al., 2008; Lohr et al., 2014).

Bauer et al (2008) illustrate the PNRCT model by including an explanatory variable.

Level 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}T_{ij} + \beta_{2j}X_{ij} + r_{ij}, \quad (2.25)$$

Level 2:

$$\beta_{0j} = \gamma_{00}, \quad (2.26)$$

$$\beta_{1j} = \gamma_{10} + u_{1j}, \quad (2.27)$$

$$\beta_{2j} = \gamma_{20}, \quad (2.28)$$

and the mixed model would be

$$Y_{ij} = \gamma_{00} + \gamma_{10}T_{ij} + \gamma_{20}X_{ij} + u_{1j}T_{ij} + r_{ij}\,, \qquad (2.29)$$

where β_{2j} is the fixed effect for the covariate and γ_{20} reflects the mean of the covariate for all subjects. The remaining coefficients were previously defined. X_{ij} is the explanatory variable (e.g., pre-scores), T_{ij} is an indicator variable, which takes the value of 1 (T_{ij} = 1) if students appear in cluster j of the treatment arm, and 0 if students appear in the control arm. Note that the treatment arm contains, again, J clusters, T_{ij} = 1 for all students for j = 1 to J, and T_{i0} = 0 for j = 0. Also note that when T_{ij} = 1, X_{ij} applies for j = 1 to J; and when T_{ij} = 0, X_{ij} applies for j = 0. In this equation, the cluster effect appears only when T_{ij} =1 for j=1 to J and does not appear when T_{ij} = 0 for j = 0. This reflects the partially nested structure of the PNRCT model adjusted by covariates.

The inclusion of an explanatory variable maintains the variance specification for the treatment and control arms as in the model from equation (1.8). The treatment arm of this model still has the within- and the between-cluster variance, while the control arm only has the between-subjects variance. Bauer et al. (2008) recommend replacing X_{ij} in the model with $(\bar{X}_{.j} + \dot{X}_{ij})$, which is a simple rescaling of the data that does not change the information in the findings. However, this re-parametrization of X_{ij} may seem impossible. The reason is that in equation (2.28), β_{2j} is specified as a constant value. Thus, there is not an $\bar{X}_{.j}$ for each cluster. In order to make this re-parametrization possible, equation (2.28) must change to $\beta_{2j} = \gamma_{20} + \bar{X}_{.j} + u_{1j}$, and $X_{ij} = \dot{X}_{ij}$. Then, substituting these terms in equation (2.25) produces equation (2.30).

$$Y_{ij} = \gamma_{00} + \gamma_{10}T_{ij} + \left(\gamma_{20} + \gamma_{21}\bar{X}_{.j}\right) + u_{1j}T_{ij} + u_{2j}\dot{X}_{ij} + r_{ij}. \qquad (2.30)$$

Covariates at level 2 are only feasible for the treatment arm, and incorporating them may happen for several reasons. For instance, Bauer et al. (2008) said that covariates might be included to explain why some groups within the treatment arm perform better than other groups. Another reason is that the variance between clusters in the treatment arm could be inflated due to the fact that the clusters are artificially formed (Lorh et al., 2014).

When including covariates in the model, especially at level 2, it is important to have enough clusters in the treatment arm to accurately determine the significance of included covariates. Additionally, it is important to measure the explanatory variable at the base line of the intervention to guarantee that it is not affected by the intervention (Lohr et al., 2014).

To exemplify the inclusion of explanatory variables at level 2, we use the following model.

Level 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}T_{ij} + r_{ij}\ , \tag{2.31}$$

Level 2:

$$\beta_{0j} = \gamma_{00}, \tag{2.32}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}W_j + \ u_{1j}. \tag{2.33}$$

The mixed model is

$$Y_{ij} = \gamma_{00} + \gamma_{10}T_{ij} + \gamma_{11}W_jT_{ij} + u_{1j}T_{ij} + r_{ij}, \tag{2.34}$$

where W_j is the level-2 variable and γ_{11} captures the effect of the interaction between the level-2 variable and the relationship between the treatment condition and the outcome variable. Again, T_{ij} is an indicator variable, which takes the value of 1 ($T_{ij} = 1$) if students appear in cluster j of the treatment arm, and 0 if students appear in the control arm. Note that the treatment arm contains J clusters, $T_{ij} = 1$ for all students for $j = 1$ to J, and $T_{i0} = 0$ for j = 0. Also note that when $T_{ij} = 1$, W_j applies for $j = 1$ to J; and when $T_{ij} = 0$, W_j does not have an effect. This is the reason the main effect on equation (2.34) is absent. This is useless for subjects in the control arm (Bauer et al., 2008). One more time, the cluster effect appears only when $T_{ij} = 1$ for $j=1$ to J and does not appear when $T_{ij} = 0$ for j = 0. This reflects the partially nested structure of the PNRCT model adjusted by covariates at level 2.

2.8. Recent Developments in PNRCT Models

In recent years, researchers have proposed and validated, by performing Monte Carlo studies, several extensions of the PNRCT model. The results of these Monte Carlo studies suggest that these models should be used. Additionally, a way to estimate effect size for the PNRCT structure has been developed. These new PNRCT models are more complex, but students are still randomly assigned into treatment and control arms, and then students in the treatment arm are clustered in groups. This section presents such models.

2.8.1. The blocked PNRCT design. The blocked PNRCT model has been recently presented for handling PNRCT data structures (Lohr et al., 2014). The main argument for this model is that sometimes randomization is done for treatment and control trials within each school or site, and since in the experiment multiple schools exist, the results can be generalized to the schools in the experiment. Thus, the main difference between this design and the pure PNRCT design is where the randomization occurs. This design allows random allocation of half of the students into the treatment arm and the other half into the control arm within each school or site.

It is worthy to note a difference between schools serving as the block units and schools serving as clusters. In the first case, student outcome scores within each school are positively correlated, which makes the estimation in the estimated treatment effects more precise. This is because students are in the same environment sharing the same conditions, which eliminates non-desirable effects coming from the school environment. On the other hand, when schools are the clusters, all students within a school are included in the treatment or control conditions, so the positive correlation between students is reduced and the estimated treatment effect would be less precise (Lohr et al., 2014). This is because students are under the influence of differential environments.

The blocked PNRCT model is specified as follows:

$$Y_{ijk} = \gamma_{000} + \gamma_{100} T_{ijk} + r_{1jk} T_{ijk} + u_{00k} + e_{ijk}, \qquad (2.35)$$

where Y_{ijk} is the test score of student i in cluster j of school k, and γ_{000} is the mean score of students in the control arm.

γ_{100} captures the treatment effect, and T_{ijk} is the treatment condition of students i in cluster j for school k (1 = treatment, 0 = control). u_{00k} is the effect in school k, r_{1jk} represents the effect of cluster j within the treatment arm in school k, and e_{ijk} represents the student residuals for students i in cluster j of school k. As in other models, this model assumes that $e_{ijk} \sim N(0, \sigma^2)$ for students in the control arm, $r_{1jk} \sim N(0, \tau_{11})$ for clusters and students in the treatment arm, and $u_{00k} \sim N(0, \tau_{\beta 00k})$ for schools.

A related model that relaxes the assumption that school effects are the same for all schools includes a term that reflects a treatment-by-school interaction. This model has the following mathematical form:

$$Y_{ijk} = \gamma_{000} + \gamma_{100} T_{ijk} + \mathrm{u}_{10\mathrm{k}} T_{ijk} + \mathrm{r}_{1\mathrm{jk}} T_{ijk} + u_{00k} + \mathrm{e}_{ijk} \, . \qquad (2.36)$$

In equation (2.36), all terms have been defined previously except for u_{10k}, which is another source of variability showing the varying effects between schools. Here, the school-level effects (u_{00k}, u_{10k}) are assumed to have bivariate normal distribution with $u_{00k} \sim N(0, \tau_{\beta 00k})$, $u_{10k} \sim N(0, \tau_{\beta 10k})$ and $Cov(u_{00k}, u_{10k}) = \tau_{\beta 00k, \beta 10k}$. This model is the same model proposed by Tesller (2014), which will be described in detail in the following section.

2.8.2. PNRCT model with three levels. Tesller (2014) proposed a three-level model as an extension of Bauer et al.'s (2008) two-level PNRCT model. Tesller's three-level model assumes that students are randomly assigned into a higher hierarchy, which is the treatment condition. These students are randomly assigned to clusters in the treatment arm or they are assigned to the control arm (no clustered structure). Contrary to the traditional PNRCT model, in the three-level PNRCT model, all students come from the

same school. Therefore, students represent level one of the model; treatment and control conditions, level two; and schools, level three.

Tesller asserted that in a PNRCT model with three levels, it is very likely that observations within schools are correlated, but not between participants within the treatment and control arms because they are randomly assigned. An additional feature of this model is that it allows for modeling the cluster conditions of subjects within schools as well as the between-school variability (Tesller, 2014). Moreover, this three-level model makes it possible to include variables at the school level, which helps to explain variability between schools. Mathematically, this model is represented by the next set of equations.

Level 1:

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk} T_{ijk} + e_{ijk} , \tag{2.37}$$

Level 2:

$$\pi_{0jk} = \beta_{00k}, \tag{2.38}$$

$$\pi_{1jk} = \beta_{10k} + r_{1jk} , \tag{2.39}$$

Level 3:

$$\beta_{00k} = \gamma_{000} + u_{00k}, \tag{2.40}$$

$$\beta_{10k} = \gamma_{100} + u_{10k}. \tag{2.41}$$

Combining all equations above produces the following mixed model:

$$Y_{ijk} = \gamma_{000} + \gamma_{100} T_{ijk} + u_{10k} T_{ijk} + r_{1jk} T_{ijk} + u_{00k} + e_{ijk}, \tag{2.42}$$

where γ_{000} reflects the average outcome of individuals in the control arm, γ_{100} captures the average treatment effect, and T_{ijk} is an indicator variable that takes a value of 1 when the subjects are in the treatment arm and zero when they are in the control arm. e_{ijk}

captures the unique individuals' random effects, r_{1jk} captures the unique clusters' random effects within the treatment arm, u_{00k} is a random effect associated with subject *i* in control condition *j* within school *k*, and u_{10k} represents a random effect for the treatment effect associated with subject *i* in the treatment condition within school *k*.

When the treatment and control conditions take the values of one and zero respectively, the model takes the following forms:

$$(Y_{ijk}|T_{ijk} = 0) = \gamma_{000} + u_{00k} + e_{ijk} \tag{2.43}$$

and

$$(Y_{ijk}|T_{ijk} = 1) = \gamma_{000} + \gamma_{100} + u_{10k} + r_{1jk} + u_{00k} + e_{ijk}. \tag{2.44}$$

Although Tesller (2014) provided no indication of the distribution of e_{ijk}, r_{1jk}, u_{00k} and u_{10k}, we could fairly assume that these effects have a normal distribution with a mean of zero and variance of σ^2_{ijk}, τ_π, $\tau_{\beta 00k}$, and $\tau_{\beta 10k}$, respectively. Therefore, $e_{ijk} \sim N(0, \sigma^2)$, $r_{1jk} \sim N(0, \tau_{11})$, $u_{00k} \sim N(0, \tau_{\beta 00k})$, and $u_{10k} \sim N(0, \tau_{\beta 10k})$.

2.8.3. PNRCT model for repeated measures. Tesller (2014) proposed another three-level model, but this time Tesller focused on individual repeated measures across time. Tesller was motivated by the fact that PNRCT data with repeated measures are frequently collected, but most researchers fail to analyze this type of data correctly. Adapting Bauer et al.'s (2008) two-level model to a three-level model, Tesller included a lower level in which measures were nested within individuals and individuals nested within treatment or control arms. Tesller's model takes the following form.

Level 1:

$$Y_{tij} = \pi_{0ij} + \pi_{1ij} Time_{tij} + e_{tij}, \tag{2.45}$$

Level 2:

$$\pi_{0ij} = \beta_{00j} + \beta_{01j}T_{ij} + r_{0ij}, \quad (2.46)$$

$$\pi_{1ij} = \beta_{10j} + \beta_{11j}T_{ij} + r_{1ij}, \quad (2.47)$$

Level 3:

$$\beta_{00j} = \gamma_{000}, \quad (2.48)$$

$$\beta_{01j} = \gamma_{010} + u_{01j}, \quad (2.49)$$

$$\beta_{10j} = \gamma_{100}, \quad (2.50)$$

$$\beta_{11j} = \gamma_{110} + u_{11j}, \quad (2.51)$$

where γ_{000} captures the average score for individuals in the control arm, γ_{010} captures the average treatment effect for individuals within the treatment arm, γ_{100} captures the average fixed slope of time for individuals within the control arm, and γ_{110} reflects the average effect of time for individuals within the treatment arm.

r_{0ij} captures the unique random effect of individuals on the average score of individuals in the control arm, u_{01j} reflects the unique random effect of individuals around the treatment effect, u_{11j} reflects the unique random effect around the average effect of time for individuals within the treatment arm, r_{1ij} reflects the unique random effect of the random slopes of the variable time in the control arm, and e_{tij} represents the score variation for the repeated measures at level 1. It is assumed that $e_{tij} \sim N(0,\sigma^2)$, $r_{0ij} \sim N(0, \tau_{\pi_0})$, $r_{1ij} \sim N(0, \tau_{\pi_1})$, $u_{01j} \sim N(0, \tau_{\beta_{01}})$, and $u_{11j} \sim N(0, \tau_{\beta_{11}})$.

Notice that it is not possible to specify a PNRCT model for a two-level repeated measure within persons at level 1 and between persons at level two. To clarify this situation, see the following model.

Level 1:

$$Y_{ti} = \pi_{0i} + \pi_{1i} Time_{ti} + e_{ti}, \quad (2.52)$$

Level 2:

$$\pi_{0i} = \beta_{00} + \beta_{01} T_{1i} + r_{0ij}, \quad (2.53)$$

$$\pi_{1i} = \beta_{10} + \beta_{11} T_{1i} + r_{1ij}, \quad (2.54)$$

In this model the treatment condition T_{1i}, which is included at level two, takes either the value of one if a person is in the treatment condition or a value of zero if a person is in the control condition. However, the model does not reflect the cluster structure of individuals. Thus modeling such a data structure, by specifying a two level repeated measures model, is not feasible. Due to the fact that in PNRCT models, individuals within the treatment arm are allocated in clusters and individuals in the control arm remain unclustered, it is necessary to have a third level to model the PNRCT data structure.

2.8.4. Partially cross-classified model. Very recently, Lou, Cappaert and Ning (2015) proposed a Partially Cross-Classified Model (PCCM). The key idea of PCCM was to model some observations that are nested within one random factor at level 1, while some other observations were cross-classified by two random factors. In order to develop a PCCM, Lou et al. used an example of evaluating the effectiveness of after-school programs. In this example, students who were not attending the after- school programs were only nested within schools, while students who attended the programs were nested within schools and within programs. Students from the same school could attend the same or different after-school programs, and students who were enrolled in any

after-school program could come from different schools. Schools and after-school programs were not necessarily nested, because the programs were community-based.

Luo et al. (2015) proposed a model for the aforementioned data structure, but they included covariates at level-1 and at level-2. This model is represented as follows.

Level 1:

$$Y_{ijk} = \beta_{0jk} + \beta_{Tjk}T_{ijk} + \sum_{p=1}^{P}\beta_{pijk}X_{pijk} + e_{ijk}, \quad (2.55)$$

Level 2:

$$\beta_{0jk} = \gamma_{00} + \sum_{q=1}^{Q}\gamma_{0q}W_{qj} + b_{00j}, \quad (2.56)$$

$$\beta_{Tjk} = \gamma_{T} + c_{00k}, \quad (2.57)$$

$$\beta_{pijk} = \gamma_{p0}. \quad (2.58)$$

Combining the previous equations, the mixed model is

$$Y_{ijk} = \gamma_{00} + \sum_{q=1}^{Q}\gamma_{0q}W_{qj} + b_{00j} + (\gamma_{T} + c_{00k})T_{ijk} + \sum_{p=1}^{P}\gamma_{p0}X_{pijk} + e_{ijk}. \quad (2.59)$$

In this model *i* represents the subjects, *j* represents one factor, and *k* the other factor at level two. Y_{ijk} is the outcome variable, X_1 to X_p represent student covariates at level one, and W_1 to W_q are covariates at level two. T_{ijk} is an indicator variable which indicates if students attend any after-school program. γ_{00} reflects the predicted outcomes for students only nested in schools when the covariates are zero, γ_q is a fixed effect capturing the impact of W_q, and $\gamma_p$captures the impact of X_p, γ_D reflects the effect of attending the after-school programs after controlling for the covariates. b_{00j} is the unique school random effect, and c_{00k} is the unique after-school program random effect. Finally, e_{ijk} reflects the level-1 residuals. It is assumed that $e_{ijk} \sim N(0, \sigma^2)$, $b_{00j} \sim N(0, \tau_{b00})$, and $c_{00k} \sim N(0, \tau_{c00})$.

To validate this model, Lou et al. (2015) conducted a simulation comparing the PCCM with other models. Their study results will be presented later in this paper.

2.8.5. Effect size for partially nested models. Accurate estimation of effect size is important. When cluster structure data exist, and it is not taken into account, the estimation of effect size can be seriously affected (Hedges, 2007, 2011). Until recently, adjusted statistics existed only for the clustering effect for two and three-level nested randomization design but not for PNRCT design (Hedges & Citkowicz, 2014). This situation was a concern for Hedges & Citkowicz (2014), so they developed adjusted statistics for accurate calculation and estimation of the size effect when PNRCT design is used.

One important issue to notice when calculating and estimating the effect size in PNRCT models is that clusters exist only in the treatment arm. Therefore, having more than one standard deviation is very likely. This fact makes it possible to have more than one definition for the mean difference between the treatment and control arms (Hedges, 2007). Hedges & Citkowicz (2014) argued that since PNRCT has clusters only in the treatment arm, two effect sizes are possible. These effect sizes are possible only if the between-cluster variation (in the treatment arm) is included when calculating or estimating the standard deviation that will be used in the effect size calculations.

The effect size can be calculated by considering whether researchers want to include the between-cluster-within-treatment standard deviation. Hedges & Citkowicz declared that this between-cluster-within-treatment variance is not taken into account when the clusters are formed because of the treatment administration.

An example of this is when clusters in an after-school program are formed because different instructors deliver a reading program to different groups of students. Under these circumstances, "one might argue that the 'natural' standard deviation of the outcome is one that does not include the extra outcome variance induced by the treatment clustering." (Hedges & Citkowicz, 2014, p. 3). Thus the effect size is calculated as follows:

$$\delta_{w} = \frac{\mu_{\cdot}^{T} - \mu^{C}}{\sigma_{W}}. \tag{2.60}$$

In other cases when the clusters already exist, Hedges and Citkowicz stated that the between-cluster-within-treatment variance has to be included as a part of the total variance. An example of this would be an experiment that includes school classrooms (here clusters already exist) in the treatment arm, but in the control arm students remain unclustered. Under this circumstance, the effect size is calculated by using the following equation:

$$\delta_{T} = \frac{\mu_{\cdot}^{T} - \mu^{C}}{\sqrt{\sigma_{B}^{2} + \sigma_{W}^{2}}}. \tag{2.61}$$

In equations (2.60) and (2.61), δ_{w} is the treatment effect size when the between-cluster variance is not taken into account and δ_{T} is the treatment effect size when the between-cluster variance is taken into account. $\mu_{\cdot}^{T}$ is the mean for the treatment arm, and μ^{C} is the mean for the control arm. σ_{W} represents the within-cluster standard deviation, σ_{B}^{2} and σ_{W}^{2} are the between and within-cluster variances respectively, and $\sigma_{B}^{2} + \sigma_{W}^{2}$ represents the total variance.

Formulas (2.60) and (2.61) are accurate when calculating the effect size for the whole population, but they bias the effect size estimation when researchers use a sample from the population. Because of this, Hedges and Citkowicz (2014) presented a set of formulas for estimating the effect size depending on whether the total standard deviation or the within-cluster standard variations are used in the estimation.

When the total standard deviation is used, the following equation is recommended:

$$d_T = \left(\frac{\bar{Y}^{T}_{..} - \bar{Y}^{C}_{.}}{S_T}\right)\sqrt{1 - \frac{(N^C + n - 2)\rho}{N-2}}, \tag{2.62}$$

and when the within-cluster standard deviation is used, the following equation should be used:

$$d_W = \frac{\bar{Y}^{T}_{..} - \bar{Y}^{C}_{.}}{S_W}, \tag{2.63}$$

where d_T is the estimated treatment effect size when S_T is used, and d_W is the estimated treatment effect size when S_W is used. S_T is the total pooled within-arm estimated standard deviation and S_W is the within-cluster standard deviation of the treatment arm. $\bar{Y}^{T}_{..}$ represents the estimated mean for the treatment arm, and $\bar{Y}^{C}_{.}$ captures the estimated mean of the control group. ρ reflects the ICC for the cluster in the treatment arm, N^C is the individual sample size in the control arm, N represents the total individual sample size, and n is the number of individuals within clusters (assuming cluster size is balanced [equal number of individuals within each cluster]) in the treatment arm.

The total pooled within-arm estimated standard deviation (S_T) comes from the total pooled within-arm estimated variance:

$$S_T^2 = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(Y_{ij}^T - \bar{Y}_{..}^T\right) + \sum_{i=1}^{N^C}\left(Y_i^C - \bar{Y}_{.}^C\right)^2}{N-2}, \quad (2.64)$$

where m is the number of clusters in the treatment arm, and *n* is the total number of observations within each cluster. Y_{ij}^T represents observation *j* within cluster *i* in the treatment arm, and Y_i^C represents observation *i* in the control arm. All other notations were previously defined.

Finally, when the cluster sizes within the treatment arm are unbalanced (unequal number of individuals within clusters), the treatment size effect is estimated as follows:

$$d_T = \left(\frac{\bar{Y}_{..}^T - \bar{Y}_{.}^C}{S_T}\right)\sqrt{1 - \frac{(N^C + \tilde{n} - 2)\rho}{N-2}}, \quad (2.65)$$

where $\tilde{n}$ is the average cluster size calculated as $\tilde{n} = (1/N^T)\sum_{i=1}^{m} n_i^2$.

2.9. Simulation Studies in PNCRT

As previously mentioned, the literature on PNRCT models is scarce. This is because the models for properly analyzing PNRCT data have only recently been developed. Currently, there are few published articles on this topic. To my best knowledge, only six simulation studies have been conducted. A summary of each study is presented next.

Baldwin et al. (2011) performed a simulation study design for evaluating models that potentially handle a PNRCT data structure. These approaches were defined as (a) ANOVA that ignores nested structure, (b) the fixed-effect approach (dummy variables for modeling cluster structure), (c) a fully-nested model assuming equal variances between arms, (d) the partially-nested model assuming equal variances between arms, and (e) the partially-nested model assuming unequal variances between treatment and control arms.

The models were evaluated under different conditions such as (a) the number of clusters (2, 4, 8 and 16 clusters), (b) cluster size (5, 15 and 30 individuals), (c) the magnitude of the ICC (0, .05, .1, .15 and .30), and (d) the degree of heteroscedasticity expressed in a ratio of variances (0.5, 1 and 2). In addition, these researchers used the between-within method (BW), the Satterthwaite method (SAT), and the Kenward-Roger (KR) method for calculating degrees of freedom for the HLMs. The models were fitted using Restricted Maximum Likelihood (RML).

To address this evaluation, Baldwin and his colleagues simulated data in such a way that the treatment effect was set to zero, the total variance for the outcome variable within clusters in the treatment arm was set to one, and the observations in the control arm were set as independent. It is worth noting that data generation and model fitting process used SAS 9.2.

The results of this research showed, in general, that PNRCT models had superior performance compared to models in which nested structure was ignored (ANOVA). Moreover, the PNRCT model (Type I error ranged from .05 to.07) performed better than the fixed-effects model (Type I error ranged from .13 to .17), and the fully-nested model (Type I error ranged from .02 to .11) in regard to Type I error rates, across the ratio of variances. Furthermore, negligible differences were found when comparing Type I error rates between the PNRCT model with equal and unequal variances (the range was from .05 to .07). Additionally, when ICC was zero, the Type I error rates decreased at a very low rate in both the ANOVA and the fixed-effect models. When HLMs were fitted with two clusters, they did not perform well compared to models with more clusters.

They showed high Type I error rates ranging from about .05 to about .35. Finally, the SAT method for estimating degrees of freedom performed better than other methods across number of clusters, cluster size, ICC values, and ratio of variances.

The results also showed that across models and simulated conditions, in the treatment arm, all models produced unbiased and reasonably efficient parameter estimates. This claim was supported by the fact that the parameter estimates had negligible bias effects (always below .02), and a Mean Square Error (MSE) which ranged from .1 to .4.

In the case of the variance components, the fully-nested and partially-nested models produced less variability (MSE ranged from .01 to .04) when the ratio between variances at level 1 was equal to .5 and one, but when the ratio of variance was equal to two and the number of clusters and size of clusters were low, the variability was high (MSE ranged from .04 to .65). However, the number of clusters and the cluster size did not have any influence on the between-cluster variance of the PNRCT models, even when heteroscedasticity was modeled. When comparing the fully-nested and the PNRCT models with equal variances across arms, the fully-nested models showed higher bias in the between-cluster variance (bias ranged from -.02 to .44), while the PNRCT models showed relatively low bias (bias ranged from -.01 to .03). Yet, the PNRCT models showed a bias in the between-cluster variance when the number of clusters, the cluster size, and the ICC were low. The rate of homogeneity, however, biased the within-cluster variance in the treatment arm (from -.23 to .45), and the between-individual variance in the control arm (from -.43 to .21).

Finally, the simulation on power rates showed a reduction when ICC increased in both the fully-nested and the PNRCT models. Nevertheless, the number of clusters and the cluster size increased the power rates. Moreover, the number of clusters was more efficient in increasing the power than the cluster size, but the difference was not remarkable. Models with equal and unequal variances produced similar power rates when the ratio of variances was equal to one.

Korendijk, Maas, Hox, and Moerbeek (2012) performed another simulation study related to the PNRCT model. These researchers attempted to determine, when analyzing PNRCT data structures, the impact of misspecifying the variance components on parameter estimates and standard errors. To achieve this goal, four models that attempted to handle PNRCT data structures were assessed: (a) Bauer et al.'s (2008) PNRCT model, (b) a fully nested model specifying control subjects as clusters of size one (NSM), (c) a fully nested model specifying control subjects as one big cluster (BCM), (d) and an analysis of variance model (ANOVA). The generation data process and the model fitting were done by using MLwiN 2.1. Additionally, the models were fitted by using REML.

These researchers generated data by varying only two specific conditions: number of clusters (10, 30, 50, and 100) and ICC values (.05, .1 and .2). In addition, one covariate at level 1 was included in the models, and its slope was set at .3. The variance in the control group was fixed at a value of 1.

Korendijk et al.'s (2012) major findings showed that the four models performed without major differences with respect to the fixed parameter estimates. On the other hand, when varying the number of clusters and the ICC, only the PNRCT model presented unbiased parameter estimates. The PNRCT model and the NSM models

estimated the fixed parameters and the standard error equally well. However, the BCM and the ANOVA models showed biased standard errors (overestimated) when associated either with the constant or the treatment effect. This condition produced inflated Type II error rates especially when the ICC and the number of clusters were large. In addition, the BCM model underestimated the level-2 variance components and overestimated the level-1 residuals. This model also inflated the treatment effect standard error estimates.

In another research study, a set of two simulation studies, were presented by Tesller (2014). In these two studies Tesller used SAS 9.2 for both simulating data and fitting models. Parameters in all models were estimated by imposing REML. In the first simulation study, Tesller extended Bauer et al.'s PNRCT two-level model to a three-level model (equations [2.37] to [2.42]). This model assumed that students were in level 1, but they were nested within either the treatment arm or the control arm at level 2, and all students (participants of the study) were nested within the same school at level 3. Tesller varied several conditions in the simulation: the ICC values at level 2 (0, .1 and .3) and at level 3 (0, .05 and .15). Additionally, the number of clusters (k), the cluster size (n), and the sample size (m) in the treatment arm were set as follows: $k = 16$, $n = 5$, $m = 70$; $k = 10$, $n = 10$, $m = 78$; and $k = 8$, $n = 15$, $m = 85$. Finally, the ratio of variances between the treatment and control arm were .5, 1 and 2.

This first study of Tesller produced a variety of findings. Here the most relevant of Tesller's results are summarized. First, the treatment effect standard error estimates were impacted by the sample size and the level 3 ICC. For instance, having few schools (e.g. 5 schools) and low values of ICC (e.g. .05) inflated the standard error of the treatment effect.

However, these standard errors presented little variability (maximum MSE = .05), suggesting efficiency in the estimation of this parameter across all simulation conditions. Note that although the standard errors presented little variation, they were still inflated. Second, the level 1 and level 2 (just for treatment at level 2) variance components, when estimated separately for treatment and control conditions, were consistently unbiased. For instance, at level 1 the average relative bias for variances was -.1% and .002% for the control and treatment arms, while the average bias for the variance at level 2 was -.3%. Third, the level-3 intercept variability estimates were somewhat biased. The mean relative bias was -.02%. However, the slopes at level 3 were inflated when the number of schools was the lowest (five schools). Fourth, when testing the model fit, the Type I error rates were below 5%. Finally, omitting level 3 of the analysis, the two-level model showed some bias in the treatment effect standard error.

In Tesller's second study, Bauer et al.'s model was extended to a three-level model but in a lower hierarchy: measures nested within students (equations [2.45] to [2.51]). To validate this model, Tesller simulated individual repeated measures across time by varying several conditions. Tesller used two sets of time points (three time points and six time points) and three ICC values (0, .05 and .15) but only at level 3. Additionally, Tesller varied *k*, *n*, and *m* in the treatment arm, which produced the following combinations: (a) $k = 16$, $n = 5$, $m = 70$; (b) $k = 10$, $n = 10$, $m = 87$; (c) $k = 8$, $n = 15$, $m = 105$; (d) $k = 32$, $n = 10$, $m = 280$; (e) $k = 20$, $n = 20$, $m = 350$, and (f) $k = 16$, $n = 30$, $m = 420$. Finally, Tesller set the ratio of variances at level 2 at 1, 2, and 4.

The major results of Tesller's second study showed that the intercept estimate bias was not relevant due to the small deviation between her results and the hypothesized parameter. However, the treatment effect bias was considered problematic because the relative bias ranged from -19% to 13%. Another finding showed that clusters and individual sample size (e.g. $k = 8$ and $n = 15$) created the largest bias effect (mean relative bias = 8.9%) on estimates. The intercept variance estimates at level 2 were biased but to a lesser extent. The slope variance estimates at level 2 in the treatment arm and the covariance estimates between intercepts and slope were biased, but mostly in the control arm (mean relative bias = 1.6%). The same pattern was found with the intercept variance estimates at level 3. However, the slope variance estimates at level 3 were biased. The relative bias ranged from -25% to 14.4%. The power rates for detecting treatment effects were impacted positively by large sample sizes and the interaction between sample sizes and ICC at level 3. Finally, Tesller found that when fitting homoscedastic models to heteroscedastic data, the standard errors of the treatment effect were not severely affected. However, at level 2, but not at level 3, the variance components were impacted. The respective effects sizes for the control and treatment arms were .13 and .15.

In a different simulation study, Sander (2011) evaluated four competing models for handling PNRCT data structure: (a) a model with pseudo-clusters in the control arm (PCM), (b) a model where the control condition was treated as one cluster (OCM), (c) a hierarchical linear model with random intercepts and (HLMRI), and (d) the PNRCT model. Sanders simulated different sets of data with different conditions.

Sanders generated different sample sizes of 40 individuals setting the following conditions: ICC values from 0 to .5 with a constant interval of .1; four effect sizes (δ), 0, .2, .5 and .8; and four numbers of clusters (2, 4, 5 and 10) with cluster sizes of 10, 5, 4 and 2 respectively. Sanders then simulated sample sizes of 160 individuals. This time the number and cluster sizes were 4, 8, 10 and 20, with respective cluster sizes of 20, 10, 8 and 4. Additionally, Sanders used three different methods for estimating the degree of freedom: (a) the BW method, (b) the SAT method and (c) the KR method. The simulation and model-fitting process were performed by using SAS 9.2. In all models Sanders used Maximum Likelihood (ML) for parameter estimates.

Sanders' major findings showed that the Type I error rate of each model relied on the estimation method, the ICC and the ratio between the number of clusters between the treatment arm and the cluster size. When the BW method was used, all models showed that the Type I error rate increased as the ICC values increased. This result differed considerably from those results achieved with the SAT and the KR methods; the models that used SAT and KR presented a conservative Type I error rate for ICC values less than .1, close to the nominal rate value for ICC values of .2 and inflated Type I error rate values when the ICC was .3.

Regarding the ratio between the treatment clusters and the cluster sample size, the results varied depending on the method for estimating the degrees of freedom. When the models included few treatment clusters with many individuals, the BW method produced a Type I error rate that was negligible in PCM and OCM models, but inflated in HLMRI and PNRCT models.

On the other hand, when the ratio between the treatment clusters and the cluster sample size was approximately one, the Type I error rate was nearly .05 for HLMRI and PNRCT models. For the SAT and KR methods, all models presented a more stable Type I error rate close to .05.

In regard to power analysis, Sanders (2011) collapsed the power rates across effect sizes, ratios of treatment clusters and treatment ICCs. Sanders found that when using the BW method, the power rate ranged from .13 to .32 across models. However, when small ratios of treatment clusters were ignored, the power rate increased in PCM and OCM models from .2 to .24 and .13 to .17, respectively. Similarly, when the KR method was used, the power rate ranged from .28 (in PCM and OCM models) to .30 (in HLMRI and PNRCT models). When ignoring small ratios of treatment clusters, negligible changes were found across models. Sanders also examined the power rate by ICC, finding that the power rates were very low (less than .2) in all models when small effect sizes were detected. The same situation occurred when examining the power rate by the ratio of treatment clusters and cluster size; the power rate performed well in all models, especially when the ratios were balanced. However, when large effect sizes were detected, the power rates were relatively low. On average, they ranged from .5 to .6 depending on the method used for estimating the degrees of freedom.

Luo, Cappaert and Ning (2015) contributed to the field of PNRCT modeling by proposing a Partially Cross-Classified Model (PCCM). To validate this model (equations [2.55] to [2.59]), Luo, Cappaert and Ning conducted two simulation studies to answer (a) how does the PCCM perform compared to a fully Cross-Classified model (FCCM) and a fully Nested Model (FNM) when the data are not proportional in size in the cross-

classified part of the data structure? and (b) under what conditions are the biases negligible to reject a PCCM and retain an FCCM or FNM? They then conducted simulation studies generating partially cross-classified data following Posner and Vandell's (1994) study for evaluating the effectiveness of after-school programs. The first simulation study presented a balanced cluster size, while in the second study the clusters were unbalanced.

Luo, Cappaert and Ning generated two subsamples in the data generation process. The first subsample included students who were only nested within schools, and the second subsample included cross-classified students in school and after-school programs. For the first subsample, these academics generated data for two covariates, one for each level of the model (level 1 and level 2). Both covariates had normal distributions with a mean of zero and a standard deviation of one. The fixed effects of the model γ_{00}, γ_{01} and γ_{10} were generated with values of 0.5 each. b_{00j} (school random effects), and ε_{ijk} (level 1 residuals) were generated assuming a normal distribution with a mean of zero each and variances equal to two and six, respectively. The ICC was set at .25 for students within schools only. The second subsample was generated with the same conditions of subsample one plus the effect of attending after-school programs (γ_{T}) and their respective random effects (c_{00k}). Thus γ_{T} was generated with a value of 0.5, while c_{00k} was generated with a mean of zero and variances of 0.89, 2.0 and 3.43.

The two simulation studies included number of schools (30 and 50), number of students per school (40), school program ratio (0.5, 1 and 1.5), percentage of crossed schools (20%, 50% and 90%), percentage of after-school program attendees per crossed

school (between 4% and 81%) and ICC values for after-school programs (.1, .2, and .3). In total, one hundred and sixty-two conditions were imposed in the simulation studies; for each condition data were generated for 1,000 individuals. This combination produced a total of 162,000 data sets. Luo, Cappaert and Ning used SAS 9.2 when generating data and fitting models. They used REML when estimating parameters.

The major findings of the first simulation study showed that the PCCM had negligible biases (close to zero) across all parameter estimates, as did the other two models. However, the other two models presented large standardized biases for the program-level variance components (from -.42 to 1.14 in the FCCM) and residual variance estimates (from -1.91 to 4 in the FCCM and from .85 to 4.26 in the FNM). When examining the bias across conditions, the program-level variance component (τ_{c00}) of the PCCM presented negligible bias in only two conditions. However, these conditions are unknown because Luo, Cappaert and Ning did not specifically mention them. The residual variance of the PCCM showed very slight bias across all conditions. The root mean square error (RMSE) of the fixed effects and the school-level variance of the PCCM were similar to those found in the other models. Their differences were less than .01.

Regarding the covariates, the coverage rate of the 95% confidence interval for the slope of the covariate at level 1 was equally consistent in the three models (acceptable range from 93.6 to 96.4%). The 95% confidence interval for the slope of the covariate at level 2 failed to be in the acceptable range in 25 conditions (the coverage rates were between 91% and 93.5%).

These coverage rates were only slightly below the acceptable range, especially for the PCCM and the FCCM models. However, this was not the case for the FNM, which performed the worst among the models. The FNM model failed to be in the acceptable range in 35 conditions. The coverage rate of the slope for the covariate capturing the after-school programs performed better for the PCCM across models. This model failed to be in the coverage range in 46 conditions. The other models failed in more than 100 conditions.

The major findings of the second simulation study showed that the fixed effect was precisely estimated across the three models. Similar patterns were found for the bias of the variance component estimates. Moreover, the three models presented an acceptable school-level variance component. In contrast, the FCCM showed a negative bias in the program-level variance component. Another notable finding is that the PCCM was more efficient than the other two models when estimating the residual variance. Regarding the covering rates of the covariates for levels 1 and 2, Luo, Cappaert and Ning obtained marked differences in this study from those in the first simulation study. The acceptable range of the coverage rates for the level-2 covariate was outside of the acceptable range in 27 (PCCM), 42 (FCCM) and 28 (FMN) conditions. The coverage rate for the after-school slopes was negatively affected by the unbalanced cluster size, but the most robust model among the three was the PCCM. The coverage rates in the PCCM were outside the acceptable range in only 11 conditions, ranging from 92.7% to 97%.

Candel and Van Breukelen (2009) assessed the impact of unbalanced clusters under the PNRCT design by using a linear mixed-effect model equivalent to the PNRCT

model. They observed the efficiency loss due to varying cluster size when estimating the variance of random effects. These researchers assumed that the data were normally distributed in both the treatment and the control arms.

Using Raudenbush and Bryk's (2002) notation, the linear mixed-effect model of Candel and Van Breukelen (2009) is

$$y_{ij} = \gamma_{00} + \left(\gamma_{10} + u_{1j} + r_{ij}^{T}\right)T_{ij} + r_{ij}^{C}\left(1 - T_{ij}\right), \tag{2.66}$$

where T_{ij} is an indicator variable (treatment =1, control = 0) that reflects the treatment condition for subject i in cluster j, γ_{10} captures the treatment effect and γ_{00} reflects the mean score for the control arm. r_{ij}^{T} is the unique random effect for individuals in the treatment arm, u_{1j} represents the unique random cluster effect also in the treatment arm, and r_{ij}^{C} is the unique random effect of individuals in the control arm. r_{ij}^{T}, r_{ij}^{C}, and u_{1j} are normally distributed with a mean of zero and variances of $\sigma^2_{\text{Treatment}}$, τ_{11} and $\sigma^2_{\text{Control}}$, respectively. This model is equivalent to the model in equation (1.8). The only difference here is that equation (2.66) differentiates the unique random effect for individuals in the treatment and in the control arms (PNRCT with heterogeneous variance). The model in equation (2.66) still differs from the standard HLM linear mixed effects model in not including the random effect for the intercept at level 2.

The simulation study conducted by Candel and Van Breukelen (2009) included unimodal, uniform, bimodal, positively and negatively skewed distributions in regard to cluster sizes. The researchers varied the average cluster size (between 6 and 10), the number of clusters (12), and the ICC values (from .01 to .30 with constant intervals of .01). Ratios of sample size between the control and treatment arms (.25, 1 and 4), ratios

of error variance between the control and the treatment arms (.5, 1 and 2), and estimation methods (ML and REML) were also varied. γ_{00} and γ_{10} were set at 50 and 5, respectively. For each simulation condition, 10,000 data sets were generated using MLwinN. The same software was used to perform parameter estimates.

The findings of this simulation study showed that ML and REML produced similar results for the relative efficiency (RE) for the parameter estimate of the treatment effect. The RE ranged from .99 to 1. Candel and Van Breukelen (2009) then presented their results only for REML. Major findings in this study indicate that the RE of the treatment effect produced extreme results in the unimodal and the bimodal distributions, but the equal cluster sizes were more efficient than the unequal cluster sizes because the RE never exceeded 1. Additionally, when the average number of subjects within the cluster was 10 and the cluster size was equal to 12, all distributions showed RE values above .92. However, the exception was the bimodal distribution, which had a value of .90. When the cluster average size was equal to six, the values were even larger than .92. Similarly, when the average cluster size of 10 was combined with the uniform distribution, the asymptotic and the simulated RE had negligible differences. Finally, when comparing the uniform distribution (cluster coefficient of variance of .27) and the unimodal distribution (cluster coefficient of variance of .42), the first distribution showed better approximation of the asymptotic and simulated RE values, but the unimodal distribution showed better approximation than the bimodal distribution (cluster coefficient of variation of .55).

With the random intercept variance, the RE value of τ_{11} was larger than 1 when the ICC was small. The unequal cluster sizes were more efficient when estimating the intercept variance. Furthermore, the simulated RE was not well described by the asymptotic RE when the ICC had small values across distributions. Finally, when the average cluster size was 10, the bimodal distribution produced values with a lower boundary of .84, but the other distributions had a lower bound of .86.

All these simulation studies validated extensions of the PNRCT model assuming that all model assumptions hold. However, real world data may produce models that violate some of the assumptions. Very little research has been conducted with respect to the robustness of the PNRCT model regarding the violation of the model assumption, especially the normal error distribution. For that reason, I argue that it is important to know the impact of departures of the level 2 error distribution from normality in the PNRCT model. The next section deals with the impact of the violation of the normality assumption. Beyond this, Chapters 3 to 5 deal with the core of this research, that is, how the PRNCRT model's outcome performs when the level 2 normal assumption does not hold.

2.10. Impact of the Violation of the Normal Distribution

Several research studies have found that the violation of the assumption of normal error term distribution is relevant in hierarchical regression. This is because when the normality assumption is violated the parameter estimates, fixed and random effects, as well as their standard errors may be seriously affected. As mentioned in Chapter 1, Raudenbush and Bryk (2002) wrote that failing to achieve the normality assumption at level 2, the fixed effects will not be biased, but hypothesis testing and confidence

intervals based on normality may be compromised, especially in the presence of outliers. On the other hand, several researchers (e.g., Shieh, 1999; Ketelsen, 2014) have found that the variance components of HLM are severely biased when the level 2 error term is non-normal.

To our best knowledge, the impact of the level 2 non-normality error on a PNRCT model's outcomes has not been studied. It may have the same implications as in HLM. However, there is no certainty about this. Although researchers have not focused on this for the PNRCT models, researchers have conducted some simulation studies on HLMs and other related models that may provide evidence of the impact that produces a violation of the normal distribution.

For instance, researchers have presented evidence of the negative impact of error heavy-tailed distributions when assessing error distributions with a high degree of kurtosis (Shieh, 1999; Shieh, Fouladi, and Pullum, 2001), Uniform distributions, Chi-square or Laplace distributions (Mass & Hox, 2004a).

Shieh (1999) conducted a simulation to evaluate the mixed effects of HLM under non-normality conditions at level 2. Shieh used a two-level model with one covariate at both level 1 and level 2, and imposed conditions such as the number of clusters (5, 20, and 80), the cluster size (5, 20, and 80), and several distributional characteristics including non-normal error distribution with different types of skewness and high values of kurtosis (heavy tails): (0, 1), (0, 3), (0, 6), (0, 25), (1, 1), (1, 3), (1, 6), (1, 25), (2, 6), (2, 25), and (3, 25). Shieh's (1999) findings suggest that with relatively few severe violations of normality the fixed parameters are robust, but when a large number of groups exist the fixed effects are even more robust. However, fixed effects standard

errors were biased. Additionally, heavy tailed distributions biased random effects, especially the variance components at level 2. The same situation occurred with their associated standard errors.

Shieh, Fouladi, and Pullum (2001) expanded Shieh's (1999) study. These researchers manipulated the error term at both level 1 and level 2 on HLM parameter estimates and included other conditions. They simulated twelve conditions for the degree of skewness (from 0 to 3) and kurtosis (from -1 to 25). Additionally, they simulated other conditions such as cluster size (n_j = 5, 20, and 80) and number of clusters (J = 5, 20 and 80). Furthermore, Shieh et al., (2001) varied the ICC (0.1, 0.3, 0.5 and 0.7) and the correlation between the level 1 intercept and slopes (r = 0.3, 0.5 and 0.7). They used a two-level model with one covariate at both level 1 and level 2. For the distribution with no skewness and heavy tails (kurtosis > 0), Shieh et al. found negligible bias on the fixed effects, but the random effects, variance-covariance components and their standard errors were negatively and severely biased.

Mass and Hox (2004a) evaluated the impact of three residual distributions (Chi-square with one degree of freedom, Uniform and Laplace distributions). In their study, they manipulated the number of clusters (J = 30, 50, 100), the cluster size (n_j = 5, 30, 50), and the intra-class correlations (ICC = 0.1, 0.2, 0.3). The model they used for performing the simulation was a two-level model with one covariate at each level. Their estimations were performed using ML and robust estimations. Mass and Hox found that the bias ranged from little bias to no bias, and the confidence intervals for the main fixed effects were not affected. In addition, they found that the standard errors of the parameter

estimates were accurate under the ML and robust estimations. They also found that the variances components (level 1 and level 2) were unbiased, but their standard errors were not always accurate. This study relied on the non-normal error distribution type.

Finally, Ketelsen (2014) examined how non-normal level 2 error distribution, among other conditions, would affect fixed parameter estimates and their standard errors. Ketelsen manipulated the ratio of imbalance, so four ratios of imbalance at level 1 were used (30:30, 23:35, 15:45, and 10:50). Furthermore, Ketelsen used three different numbers of clusters (600, 900, and 1,500), five ICC values (.05, .10, .15, .20, .25), and three level 2 error distributions. The level 2 error distributions were adjusted to two different degrees of skewness (0, 1.63 and 2.82) and kurtosis (0, 7.0 and 15.0). Three level-2 error distributions emerged from the combination of skewness and kurtosis (0:0, 1.633: 7.0, and 2.828:15.0). All conditions were simulated using a two-level model.

Ketelsen's findings suggest that non-normality error distribution at level 2 biased the standard error of fixed effects at level 1 and the variance components. Both standard deviation and standard errors became inflated when the error distribution had extremely non-normal conditions. Ketelsen also determined that under extreme conditions of non-normality, the Type I error rate was affected, but power was not affected under any type of non-normality.

In summary, these studies have showed that non-normal error distribution does not seriously affect the quality of the fixed parameter estimates. However, the standard error of the fixed effects may be found to be biased, depending on the non-normal error distribution type. Additionally, non-normal error distribution may bias random effects,

especially the variance components at level 2. Non-normal error distribution with a high degree of kurtosis (Shieh,1999; Shieh, Fouladi, and Pullum, 2001) biases the level 2 variance component; but distributions such as Uniform, Chi-square, and Laplace may not bias the level 2 variance component.

However, regardless of the non-normal error distribution type, the variance standard errors have been found to be biased (Mass and Hox, 2004a; Shieh,1999; Shieh, Fouladi, and Pullum, 2001). The confidence interval of the fixed effects has also been affected under non-normality. Finally, the lack of normality may negatively affect the Type I error rate, but may not impact the power of the fixed effects (Ketelsen, 2014).

Although the previous literature shows the impact of violating the level 2 error distribution on HLM outcomes, it is not possible to automatically generalize these results to the PNRCT model. This is because the violation of the level 2 error distribution, in HLM, has been simulated based on the random intercepts, which the PNRCT model does not have. This fact may produce different results, compared to the HLM, on the PNRCT model when the normality assumption does not hold at level 2. Since no research has been performed on the PNRCT model, measuring the impact on non-normal error distribution at level 2, this is an issue that needs to be studied.

2.11. Chapter Summary

First, researchers have proposed some models for handling the nested structure of the PNRCT design. Models such as standard regression analysis and ANOVA ignore the cluster structure of the PNRCT design. On the other hand, models such as clusters as fixed effects, clusters with one individual in a control condition, pseudo-clusters in a

control condition, and one large cluster, take into account the clusters structure of the PNRCT. However, the evidence has showed that these models have failed to properly model the PNRCT data structure.

Second, Bauer et al. (2008), proposed a model (the PNRCT model) that properly takes into account the nested data structure of the PNRCT design. The PNRCT model does not allow random intercepts; however, it is still possible to calculate the ICC, because the treatment condition presents a nested data structure. One important property of the PNRCT model is that it can handle unequal variances in the treatment and control conditions.

Third, covariates can be included at levels 1 and 2 of the PNCRT model, which makes it possible for researchers to use the model in quasi-experimental and correlational studies. As in the HLMs, in the PNRCT models the covariates can be centered.

Fourth, more recently, researchers have extended the "pure" PNRCT model to more complex models. These models are (a) the block PNRCT model, (b) the PNRCT model with three levels, (c) the PRNRCT model for repeated measures, and (d) the Partially Cross-Classified model. In addition, some formulas have been proposed to estimate the effect size in the PNRCT model.

Fifth, a few Monte Carlos studies (six studies) have been performed in order to evaluate and validate the PNRCT models previously mentioned. Among these studies, only two studies evaluate the adequacy of the "pure" PNRCT" model. When performing these Monte Carlo studies, researchers have included several conditions such as levels of ICC, number of clusters, within-clusters observations, ratios of sample size, ratios of variances, and effect size. Other types of conditions that have been studied are the

methods of estimation (REML and FML), and the methods of estimating the degrees of freedom (BW, SAT, and KR).

Finally, all Monte Carlo studies performed on the PNRCT models have assumed that the models' assumptions hold. However, this may not be true in real world data. Thus, this chapter presented evidence that considers the consequence when the error normality assumption is violated in HLMs. It is important to remember that this issue has not been addressed in PNRCT models, but since PNRCT models includes a treatment with HLM structure, the evidence may be related to PNRCT models. In general, when the assumption of normal error distribution is violated, the fixed effects remain untouched but their standard and the random effects are biased. Relatedly, some evidence points out that the Type I error rate and power may be affected when the normality assumption is violated.

Chapter 3

Methods

This study focused on the situation in which the assumption of normality of residuals at level 2 of the PNRCT model does not hold, specifically when non-normal error distributions are heavy tailed (t distributions with four and eleven degrees of freedom). Thus, the present study examined the effect of violating the assumption of normality at level 2 in the PNRCT model on the sensitivity of parameter estimates (fixed effects), Type I error rate, and power in the "pure" PNRCT model and in a PNRCT model adjusted by one covariate at level 1. The goal of this study was to provide educational researchers a basis on which to judge whether the conditions for PNRCT analysis are plausible. Additionally, researchers may be benefited from information regarding the impact of departures from the underlying level 2 normality assumption. Furthermore, the study was designed to increase the literature about PNRCT by investigating the performance of PNRCT models under specified conditions.

This chapter includes several sections describing the research design of the present study. The first section presents the research question of the study. Then the simulated models are introduced in section two. The third section introduces the parameters used in the study. Section four and five discuss the independent (conditions of the study) and the dependent variables. Section six and section seven describe the data generation procedures and the number of replications, respectively. Finally, in the last two sections, the methods of estimation and the data analysis are presented.

3.1. Research Questions

The following research questions were explored in the present simulation study: How is PNRCT estimation (fixed effects: treatment effect and covariate effect) and inference (Type I error rates and power) influenced by

1. Cluster size?
2. Number of clusters?
3. Levels of the intra-class correlation coefficient (ICC)?
4. Heavy-tailed error distributions at level 2 (a t distribution with four degrees of freedom and a t distribution with eleven degrees of freedom)?

Each research question is important but the last one is most important because of the previously documented gap in the literature examining the impact of non-normal error distributions at level 2.

The simulation data matrix represented cross sectional data and it was simulated and analyzed by using the lme4 and lmerTest R packages (Bates, Maechler, Bolker, & Walker, 2015; Kuznetsova, Brockhoff, & Christensen, 2015; R Core Team, 2015).

3.2. Simulated Model

The present study utilized two PNRCT models for cross-sectional data: the "pure" PNRCT model (no covariates included, hereafter the "Model A") and a PNRCT model including one covariate (hereafter the "Model B").

3.2.1. Model A. The model specified for the current simulation study with no covariates is regarded as the "pure" PNRCT model, and it is used in the process of generating data. The model was previously specified in equations (1.8).

3.2.2. Model B. This model is a PNRCT adjusted by one covariate at level 1. The rationale for using a model with one covariate is that this may improve the precision of the treatment estimates under certain conditions, especially when the covariate predicts the outcome, but the covariate does not have to account for too much variance in the outcome variable. Thus, Model B was created by introducing a covariate at level 1. This model is the same as in equation (2.29).

3.3. Parameters

Researchers have used a variety of parameter values when performing simulation studies. When possible, researchers take these parameters from applied settings (e.g. Maeda 2007), and sometimes parameters are specified in such a way as to facilitate the simulation process. Two examples of this last situation are Baldwin et al. (2011) and Korendijk et al. (2012).

Baldwin et al. (2011) set γ_{00}, the mean of the control group, and γ_{10}, the treatment effect, equal to zero but did not set any specific value for σ^2, the residual variance, or τ_{11} (the level 2 variance); rather, they set $\sigma^2 + \tau_{11} = 1$. Korendijk et al. (2012) set the mean of the control group (γ_{00}) as a value of 1, the treatment effect (γ_{10}) as 0.3, the covariate effect (γ_{10}) as 0.3, and the within group variance (σ^2) in the treatment group as 1. Thus, τ_{11} was a function of the ICC.

Two sets of parameter values were used in this study. The first set of parameter values were zero for γ_{00} (mean of the control group) and γ_{10} (treatment effect) in Model A and Model B, and zero for γ_{20} (covariate effect) in Model B. This first set of values indicated that the null hypothesis was correct and allowed determination of the impact on the Type I error rate (See Section 3.5, Subsection 3.5.3).

The second set of parameter values were a value of zero for γ_{00} (mean of control group) and a value of one for γ_{10} (treatment effect) used for Model A and Model B. In addition, for Model B, a value of one was also used for γ_{20} (covariate effect). The second set occurred when the null hypothesis was false and allowed determination of the impact on power. Additionally, all PNRCT model's outcomes were examined when these values were used in simulating the data. The rationale for using these standardized values was to keep the simulation process as simple as possible and model realistic data conditions. In addition, the metric of the data remained small and easy to interpret.

The values of τ_{11} were different across level 2 error distributions in both Model A and Model B. For the normal distribution, τ_{11} was set to 1; for the t distribution with four degrees of freedom and for the t distribution with eleven degrees of freedom τ_{11} was 2 and 1.22. This was due to the fact that the variance for these distributions is estimated by $\nu/ \nu\text{-}2$, where ν stands for degrees of freedom. Although the level 1 error distribution remained normal across the level 2 error distributions, it was impossible to keep the same values for σ^2. This happened because the values of the ICC were controlled in the experiment, and because σ^2, ICC and τ_{11} are related to each other by equation (2.20) which is $ICC_{treatment} = \frac{\tau_{11}}{\tau_{11}+\sigma^2}$. By solving this equation for σ^2, equation (3.1) is produced.

$$\sigma^2 = \frac{\tau_{11}}{ICC_{treatment}} - \tau_{11} \, . \qquad (3.1)$$

Note that for either equation (2.20) or equation (3.1) it was not possible to control the three parameters at the same time but rather was possible to control only two of them. Since in this study, the ICC and τ_{11} were controlled, σ^2 has to be calculated by equation

(3.1). As a result it was not possible for τ_{11} and σ^2 to be the same across level 2 error distribution. Similar decisions about what values to hold constant and what values to vary in a simulation can be found in most simulation studies of these models (e.g., Baldwin et al., 2011; Korendijk et al., 2012).

The set of values for σ^2, τ_{11}, and ICC used in the present study are presented in the following table.

Table 1

Values of σ^2 and τ_{11} by ICC Level

	ICC	τ_{11}	σ^2
u_{1j} ~ N(0,1)	0.05	1.000	19.000
	0.15	1.000	5.667
	0.25	1.000	3.000
u_{1j} ~ t (0, 1.414)	0.05	2.000	38.000
	0.15	2.000	11.333
	0.25	2.000	6.000
u_{1j} ~ t (0, 1.106)	0.05	1.222	23.222
	0.15	1.222	6.926
	0.25	1.222	3.667

In practice, when τ_{11} increases, ICC also increases, showing more variability in the between-group variance. In the case of this study, it did not happen because τ_{11} was fixed to the same value across the levels of ICC[2]. It is important to notice that this last situation may be unrealistic for normal distributed data. The simulation restrictions then affect the interpretation of effects due to ICC, because for this study the ICC is really the reduction of within-cluster variance, not the increase of between-cluster variance.

This meant that Spybrook et al.'s (2011) key idea on the inverse relationship of

[2] Hereafter ICC is the same as $ICC_{treatment}$.

ICC and power did not hold. Spybrook et al. (2011) wrote that as ICC increases, power decreases for fixed values of the sample size and the number of clusters. This happened because Spybrook et al. constrained the total variance[3] ($\tau_{11} + \sigma^2$) to 1. Thus algebraic manipulation of equation (2.20) reveals that the ICC = τ_{11} and 1-ICC = σ^2. Under these conditions, as ICC levels increase, τ_{11} also increases, which implies a larger proportion of between-cluster variance. This directly impacts the standard error of fixed effects (e.g., $\hat{\gamma}_{10}$), which is

$$SE_{(\hat{\gamma}_{01})} = \sqrt{\frac{\left(\tau_{11} + {}^{4\sigma^2}/_{\mathrm{n}}\right)}{\mathrm{J}}} \quad , \tag{3.2}$$

where n is the total sample size and J the number of clusters. In this equation, as the ICC increases τ_{11} also increases but σ^2 decreases, which causes the standard error of the fixed effect also increases. When the standard error increases power decreases.

This situation is easy to see after algebraically manipulating equation (3.2) to obtain

$$SE_{(\hat{\gamma}_{10})} = \sqrt{\frac{\left(ICC + {}^{4(1-ICC)}/_{\mathrm{n}}\right)}{\mathrm{J}}} \tag{3.3}$$

This equation shows that as ICC increases when n and J are fixed, the standard error increases, thus power decreases.

In the present study τ_{11} was fixed across the ICC level for each distribution (See Table 1). Note that it was not assumed $\tau_{11} + \sigma^2 = 1$; rather $\tau_{11} = 1$ was assumed for the level 2 normal error distribution and τ_{11} was 2 and 1.222 for the t distributions with 4 and 11 degrees of freedom. In equation (3.2), τ_{11}, n and J are fixed but σ^2 decreases as the

[3] Spybrook et al. (2011) were referring τ_{00} and σ^2 but this also applies for τ_{11} and σ^2.

ICC increases (See Table 1). Therefore, the standard error decreases as the ICC increases and when the standard error decreases power increases.

3.4. Independent Variables

A 3 × 3 × 3 × 3 factorial design was used in the present simulation study. The following independent variables were used: (a) Cluster size, (b) Number of clusters, (c) ICC values, and (d) Distribution type of the level 2 error distributions. In total, this study had 81 conditions. Cluster sizes were 6, 17, and 32 and level 2 sample sizes were 10, 30, and 50 clusters. The ICC factor had levels of .05, .15, and .25, and error distributions were a normal distribution, a *t* distribution with four degrees of freedom, and a *t* distribution with 11 degrees of freedom. Rationales for these choices appear next.

3.4.1. Cluster size. Kreft (1996) proposed as a rule of thumb to use 30 subjects for the cluster size when performing hierarchical linear modeling (HLM). However, in applied settings, the number of subjects (e.g., students) in the cluster size may vary from a small number (e.g., 1, 2, 3, or 4) to a large number (e.g., over 50). In PNRCT models, it is hard to determine frequently used cluster sizes due to the lack of studies using this model in applied settings (Sanders, 2011). Some examples of PNRCT studies include the following.

Savage, Abrami, Hipps and Deault (2009) conducted a study in which students were taught in groups of four within the treatment condition. Another example is the research study conducted by Roberts et al., (2011), in which subjects were grouped within clusters of four. Bauer et al. (2008) illustrated the use of PNRCT by reanalyzing the data from the Reconnecting Youth Program (RY), a preventative intervention

program. In the intervention condition, the clusters had 5 to 15 subjects. Baldwin et al. (2011) also illustrated the use of PNRCT models by analyzing data from the Body Project (BP). The clusters in the treatment condition had an average of 6.5 individuals within clusters. Monte Carlo studies conducted with PNRCT models have used a relatively small number of subjects in each cluster. For example, Sander (2011) used 2, 4, 5, and 10 observations in each cluster; Baldwin et al. (2011) used 5, 15, and 30; Teslle r (2014) used 5, 10, and 15; and Korendijk et al. (2012) and Luo et al. (2015) kept the number of individuals across clusters constant, 5 and 40. Finally, Candel and Van Breukelen (2009) used an average of between 6 and 10 in a cluster.

The above examples illustrate that the cluster size in a PNRCT model has generally been small. Acknowledging this, in the present study three values were randomly selected. To perform this selection, a random generation syntax was written in R software in such a way that an upper and lower boundary of 2 and 40 was set (this range was selected from the studies previously reported), which allowed three random numbers to be obtained in one draw. These numbers were 6, 17 and 32.

3.4.2. Number of clusters. Many researchers have agreed that the number of clusters is more important than the cluster size (Mass & Hox, 2005; Paccagnella, 2011; Van der Leeden & Busing, 1994; Van der Leeden, Busing, and Meijer, 1997). This suggests that selecting the number of clusters is a critical decision. However, there is little agreement regarding the required level 2 sample size. The spectrum of recommendations ranges from six to 100. For instance, Browne and Draper (2000) recommended between six and twelve for variance estimates, and Mass and Hox (2004b) stated that ten clusters

are enough for estimation of fixed effects, 30 for contextual effects and 50 for standard error estimates. Busing (1993) has recommended 100 in order to have valid estimates.

The number of clusters found in simulation studies of PNRCT models ranges from two (Baldwin et al. 2011; Sander, 2011) to 100 clusters (Korendijk et al., 2012), which may be impossible to find in applied settings. Table 2 shows in detail the information regarding the number of clusters for the PNRCT simulation studies.

Table 2

Summary of Level 2 Sample Sizes in PNRCT Simulation Studies

Study	Sample size at level 2 (number of clusters)
Baldwin et al. (2011)	2, 4, 8, 16
Korendijk, Maas, Hox, and Moerbeek (2012)	10, 30, 50, and 100
Tesller (2014)	8, 10, 16
Sander (2011) First study	2, 4, 5 and 10
Sander (2011) Second study	4, 8, 10, 20
Luo, Cappaert and Ning (2015)	30 and 50
Candel and Van Breukelen (2009)	12

The above information raises the question of the accuracy of estimation, especially when the sample size is small. This is particularly important because the level-2 sample sizes by definition are smaller than at level 1. Therefore, for the purpose of this study, three level-2 samples sizes were selected: 10, 30 and 50 clusters.

These values were selected because they have been recommended as plausible numbers of clusters. As mentioned previously, Browne and Draper (2000) recommended between six and 12, and Mass and Hox (2004b) asserted that ten clusters are enough for estimation of fixed effects, which is the focus of the study. Therefore, this study

examined whether Browne and Draper's and Mass and Hox recommendations (here, 10 clusters) was plausible under normality and non-normality error distributions. Furthermore, this study examines 30 and 50 clusters as suggested by Mass and Hox (2004a). This is because 30 clusters showed unstable results in Maeda's (2007) study and because 50 clusters are frequently found in school research (Mass & Hox, 2005a).

3.4.3. Intra-class correlation coefficient. The intra-class correlation coefficient (ICC) is an important factor in the design and analysis of PNRCT because it represents the variation between clusters in the treatment group. If the ICC is not taken into account, when in fact it is present, the assumption of independent observations is violated (Ketelsen, 2014). The effects of ignoring the ICC are well documented; this underestimates standard errors, increases the Type I error rate, and produces an inflation of R^2 (Osborne, 2000; Raudenbush & Bryk, 2002; Singer, 1987).

The range of ICC values, across schools, in applied settings is quite large. For instance, Zopluoglu (2013) found in large-scale studies (Programme for International Students Assessments [PISA] and Trends in International Mathematics and Science Study [TIMSS]) an average of ICC values across years (1995, 2001, 2003, 2006 and 2007) and across domains (Mathematics, Reading, and Science) from .19 to .31. Another example is Hedges and Hedberg (2007) who reported average values from .05 to .15, across schools, in large-scale studies. Several researchers (e.g., Bloom, Bos, & Lee, 1999; Bloom, Richburg-Hayes, & Black, 2007; Hedges & Hedberg, 2007; Schochet, 2008) have reported a range between .15 and .25 in achievement data in the USA.

In simulation studies, researchers have used values that overlap with values in applied settings. For instance, Konstantopoulus (2009) used ICC values of .1 and .2. In addition, Ketelsen (2014) reported values from .05 to .25. For PNRCT models, researchers have used a variety of ICC values, ranging from 0 up to .5. Table 3 summarizes these values.

Table 3

Summary of Intra-Class Correlation in PNRCT Simulation Studies

Study	Intra-Class Correlation (ICC)
Baldwin et al. (2011)	0, .05, .1, .15 and .30
Korendijk, Maas, Hox, & Moerbeek (2012)	.05, .1 and .2
Tesller (2014)	0, .1 and .3
Sander (2011) First study	0, .1, .2, .3, .4, and .5
Sander (2011) Second study	0, .1, .2, .3, .4, and .5
Luo, Cappaert and Ning (2015)	.1, .2, .25 and .3
Candel and Van Breukelen (2009)	from .01 to .30 with constant intervals of .01

Unfortunately, no PNRCT model simulation research has studied the effect of ICC in the presence of non-normality. However, some evidence from HLM illustrates several problems in parameter estimates, Type I error rate, and power. For instance, Mass & Hox (2004a) found that when level 2 error distributions are non-normal and values of ICC are 0.1, parameter estimates have a statistically significant bias. Shieh, Fouladi and Pullum (2001) additionally found that an ICC of 0.5 increases the relative bias of fixed parameter estimates in the presence of non-normally distributed errors. Ketelsen (2015) also found weak evidence that the interaction between ICC values and non-normal error distribution may affect relative bias. Moreover, Ketelsen found strong

evidence that the RMSE is biased in the presence of non-normality at various values of the ICC.

Regarding the impact of the ICC on Type I error rates of parameter estimates, Ketelsen found that as the ICC increases (.05 to .25) the Type I error rate also increases over the nominal rate. In addition, in Ketelsen's research study, there was weak evidence that a relatively high value of the ICC (.25) and a small number of clusters (20) may reduce the power below .80. However, Ketelsen did not discuss this finding.

Finally, Shieh, Fouladi and Pullum (2001) found that the ICC values (.1, .3, .5 and .7) downwardly bias the random effects at level 2 in the presence of non-normality, and Mass and Hox (2004a) determined that the coverage of confidence intervals of random effects are negatively affected by high ICC values.

The above evidence indicated that the ICC may have a negative impact on parameter estimates, Type I error rate and power. In light of this situation, this study examines three levels of ICC: .05, .15, and .25. Although Table 2 shows values of ICC higher than .25, the highest level of ICC was .25 because this is the highest value of ICC in the U.S. achievement data. I believe that the ICC values of the present study fairly represent the range of values in Table 3.

3.4.4. Non-normal error distributions. Micceri (1989) showed that real world data hardly ever present a normal distribution. When using data from applied settings, the error distribution from regression models can produce any distributional form. When this happens, the model assumption of normality is violated, affecting the quality of parameter estimates and leading researchers to incorrect conclusions.

Non-normal error distribution may produce a heavy-tailed distribution. This may be the reason some simulation studies regarding non-normal assumption issues analyze heavy-tailed error distributions (Ketelsen, 2014; Shieh, 1999; Shieh et al., 2001). Several situations can cause a heavy-tailed distribution to appear in applied settings. One of these situations is when the distribution of an outcome variable includes a large proportion of two subpopulations located in the upper and lower tail of the distribution. For instance, it is well known that SES is related to student achievement. Low SES students tend to perform poorly while high SES students tend to perform well in achievement tests. If for some (unknown or known) reason a large proportion of very high and very low SES students are in the sample, the distribution of student achievement will likely be heavy-tailed. A heavy-tailed distribution may also appear because the distribution of an outcome variable has a large percentage of individuals in one tail of the distribution. For instance, two percent of school districts in a city have 60% of student absences during a year.

The present research includes the normal error distribution and two symmetrical heavy-tailed error distributions. These heavy-tailed error distributions are two t distributions with four and 11 degrees of freedom, respectively. To explain this clearly, these distributions are presented in the following figure.

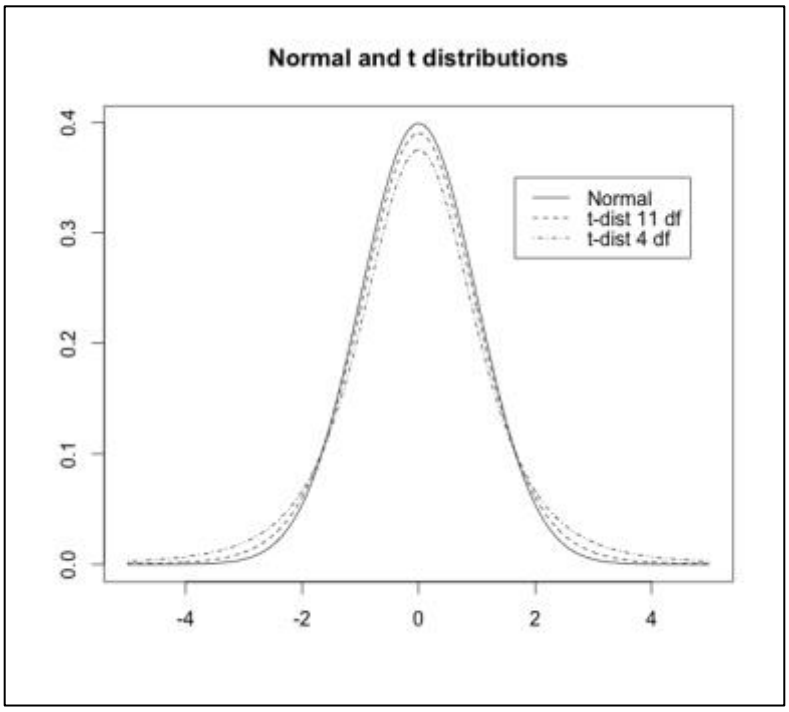

Figure 1. Normal distribution, t distribution with four degrees of freedom, and t distribution with 11 degrees of freedom.

The probability density function of a t distribution is represented as

$$f(t,\ \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \tag{3.4.}$$

where ν is the number of degrees of freedom and Γ is the gamma function.

As shown in Figure 1, t distributions with four and eleven degrees of freedom are more heavy-tailed than the normal distribution. Hogg (1974) introduced two tail weight measures that seem to be proper for symmetrical distributions such as t distributions. These measures are known as Q and Q_1, and they are defined as

$$Q = [U(.05) - L(.05)/U(.50) - L(.50)] \tag{3.5}$$

and

$$Q_1 = [U(.20) - L(.20)/U(.50) - L(.50)]. \tag{3.6}$$

In equations (3.5) and (3.6), U(β) is the average of the largest $n\beta$ order statistics and L(β), which uses the smallest item, has a similar definition. These two measures have been used in applied settings to determine whether a data distribution is heavy-tailed (Micceri, 1989).

To illustrate the use of these measures, the value of Q was estimated by simulating a set of 50 replications (in R software) for the previously mentioned t distribution and normal distribution. Within each replication, 60 observations were generated. These observations were equivalent to 10 clusters and six observations within each cluster. Q was calculated by using the "npsm" package (Kloke & McKean, 2014) in R software. In addition, p-values below and above -3 and 3 studentized values for both a t distribution with four degrees of freedom and a t distribution with eleven degrees of freedom were calculated. These p-values were then compared with a p-value below and above -3 and 3 standardized values from the normal distribution. The results are presented in Table 4.

Table 4

Q and p-values Comparison for the Distributions: Normal, t with Four Degrees of Freedom, and t with 11 Degrees of Freedom

Distribution	Q	p-value
Normal	2.552	0.004
t distribution 4 df	2.655	0.045
t distribution 11 df	3.273	0.011

Note: The p-values in the table show collapsed probability for above 3 and below -3 studentized and standardized values.

The Q values for the normal and t distributions with four and 11 degrees of freedom show that the t distribution has heavier tails than the normal distribution. The p-values also show that the t distributions with four and 11 degrees of freedom have higher p-values than the normal distribution in the tails, which implies heavier tails than the normal distribution.

In this study, the t distributions were used as error distributions for three reasons. First, these distributions may have the potential to negatively affect parameter estimates, Type I error rates and power. Second, Raudenbush and Bryk (2002) wrote that when fitting HLMs with educational data, heavy-tailed error distributions may negatively impact the outcomes of HLMs. When this situation happens, the fixed effects will not be biased, but hypothesis testing and confidence intervals based on normality may be compromised, especially in the presence of outliers. Finally, in the applied settings of education and psychology, heavy-tailed distributions, of which t distributions are a part, are more common than expected (Micceri, 1989).

In addition, the t distribution is symmetric and very similar to the normal distribution. Thus if a t error distribution at level 2 negatively impacts the PNRCT model's outcomes, there is no reason to think that other heavy-tailed distributions will not produce negative impacts on the PNRCT model's outcomes. If no negative impacts are found, further studies may examine other heavy-tailed distributions.

3.5. Dependent Variables

In general, the dependent variables used in simulation studies are related to the accuracy (sensitivity of estimators) and dispersion of the fixed effects.

In addition, Type I error and power are often assessed (Ketelsen 2014; LeBeau, 2013). For each cell of the factorial design, the sensitivity of estimator measures was calculated as in similar simulation studies on PNRCT and HLM models (Ketelsen, 2014; Kim, 1990; Korendijk et al., 2012; LeBeau, 2013; Maeda, 2007; Sanders, 2011; Tesller, 2014). The accuracy measures, Type I error rate, and power are discussed in the following subsections.

3.5.1. Sensitivity of estimators. As previously mentioned, a heavy-tailed error distribution may affect fixed effects. It was then necessary to examine to what degree an estimator parameter recovered the parameter values under the established conditions in the study. Thus the parameter estimates were collected and evaluated across simulated conditions in terms of Relative Bias (RB), and Root Mean of Square Error (RMSE) for both Model A and Model B.

RB is an indicator of the magnitude of observed bias, which often is interpreted as the percentage of bias (under or overestimated) of a parameter. One advantage that *RB* has is that this permits comparison of the amount of bias among parameters that differ in magnitude (Krull, 1997). Furthermore, this index takes into account the direction of the bias (Ketelsen, 2014, Maeda, 2007; Sanders, 2011; Tesller, 2014). This measure is calculated as a percentage of the true parameter value, and it is defined as the difference between the true parameter value and a parameter estimate, divided by the true parameter value. Consequently, RB is scaled by the true parameter value. The following equation represents the RB:

$$RB = \frac{\hat{\bar{\theta}} - \theta}{\theta}, \tag{3.7}$$

where $\hat{\bar{\theta}}$ is the average observed estimated parameter and θ is the true parameter value specified for the study. Since each cell of the design produced one value of RB, the results were analyzed quantitatively.

The literature does not report a rule of thumb for assessing the relative bias in Monte Carlo studies. Researchers frequently use their own judgment, considering the extent to which the parameter estimates deviate from the true parameter value without any referent value (e.g., Shieh et al., 2001). Other researchers have defined arbitrary cutoff criteria to assess the relative bias. For instance, Shieh and Fouldai (2003) and Delpish (2006) define, without any further explanation, small to negligible bias as when the relative bias is less than 5%; moderate or medium bias as when the relative bias is 5% to 20%; and large bias as when the relative bias is more than 20%. Other researchers have used more restrictive values. For instance, Maeda (2007) and Ketelsen (2014) claim that relative bias exists when the parameter estimates depart more than .001 from the true parameter.

I argue that values such as 0.1% (.001) or 20% are very conservative and very liberal, and that using these could lead to extreme results (e.g., all parameter estimates with bias, or no bias in any parameter estimate). I believe that in models such as PNRCT and PNRCT adjusted by the use of a covariate, 5% is reasonable to determine the existence of bias. The reason is that PNRCT and PNRCT adjusted, used in this research, are very parsimonious models, and the parameter estimation is not so complex. Thus, it was not necessary to be very conservative nor very liberal. Therefore, in this study a criterion of 5% ($\pm$.05) was used to evaluate the relative bias of the fixed effects. This

meant that if the relative bias was within this range, the relative bias was not considered to be a problem.

The RMSE measured the variability of the parameter estimates produced by each replication in the study. In other words, it measured the average distance of a parameter estimate from the true parameter value. RMSE is a measure of accuracy, which is defined as the overall distance between the observed values and the true value (Bainbridge 1985, Zar 1996, Jones 1997, Krebs 1999). This measure was affected by both bias and the dispersion of estimates. The RMSE indicated good performance of a parameter estimate when its values were small. The RMSE was estimated by obtaining the square root of the summation of all the square deviations between the parameter estimates and the true parameter value, divided by the number of replications. The following equation was used to calculate the RMSE:

$$RMSE = \sqrt{\sum_{1}^{NR} \frac{\left(\hat{\hat{\theta}}_i - \theta\right)^2}{NR}}, \tag{3.8}$$

where RMSE was the root mean square of the error. $\hat{\hat{\theta}}_i$ and θ were previously defined. *NR* is the number of replications.

3.5.2. Power. Power is the probability of correctly rejecting a false null hypothesis. Power is usually evaluated as the proportion of the number of rejected (false) null hypotheses for a test of a parameter at an alpha level of .05. Power was evaluated for γ_{10} (treatment effect) for both Model A and Model B, and γ_{20} (covariate effect) was evaluated for Model B. Power was evaluated only when the set of parameter values were equal to 1 for γ_{10} and γ_{20}. Thus, by having a parameter estimate equal to 1, the false null

hypothesis would be, for instance, $\gamma_{10} = 0$, so every time this hypothesis was rejected, the false null hypothesis was being rejected correctly. Power then was evaluated by using the following equation:

$$PSP = \frac{nsp}{NR}, \tag{3.9}$$

where *PSP* stood for the proportion of significant statistical tests, and *nsp* was the number of significant parameter estimates at the specified alpha level. *NR* has already been defined. Then, *PSP* was compared with a cutoff of .80. Significant departures from this cutoff were considered low and high, respectively.

3.5.3. Type I error rate. The Type I error rate is present in a test when a true null hypothesis is incorrectly rejected. Typically, the Type I error rate in Monte Carlo studies is evaluated as the rate of the proportion of rejections of a true null hypothesis. In this study, the Type I error rate was evaluated for the treatment effect (γ_{10}) for both Model A and Model B, and the covariate effect (γ_{20}) was evaluated for Model B. The Type I error rate was evaluated only when the set of parameter values was equal to zero. By having a parameter estimate equal to zero, the true null hypothesis was, for instance, $\gamma_{10} = 0$, so every time this hypothesis was rejected, I knew that the true null hypothesis was being rejected incorrectly. The following formula was used to evaluate the Type I error rate:

$$Type\ I\ error\ rate = \frac{NPVLA}{NR}, \tag{3.10}$$

where *NPVLA* is the number of significant statistical tests. *NR* was previously defined.

Sometimes, it is desirable to construct a confidence interval around the specified alpha value (Type I error rate). Proportions within this interval are not considered inflated values. However, proportions below the lower limit and above the upper limit are

considered conservative or inflated, respectively (Harwell, 2015). Following Harwell (2015), the confidence interval for this study was constructed by $0.05 \pm 1.96\sqrt{0.05(1-0.05)/1000}$. The interval then ranges from .036 to .063. Notice that the number of replications is 1000. Section 3.7 explains how this number of replications was reached.

3.6. Data Generation Procedure

In the present study, several steps were followed for simulating the data. These steps were similar to those used in several previous HLM or PNRCT model studies (e.g., Baldwin et al. 2011; Coleman, 2006; Darandari, 2004; Ketelsen, 2014; Maeda, 2007). The level 1 and level 2 residuals were simulated separately from known distributions, u_{1j} was simulated assuming the aforementioned distribution (a normal distribution, a *t* distribution with four and a t distribution with 11 degrees of freedom), and r_{ij} was simulated from a normal distribution using the values in Table 1. The Y_{ij} values were created by adding up all elements in equation (1.8) and equation (2.29) for Model A and Model B. This implied T_{ij}, X_{ij}, u_{1j} and r_{ij} had to be simulated. T_{ij} was generated as an indicator variable and X_{ij} was sampled from a continuous normal distribution with a mean value of zero and a standard deviation of one.

It is important to notice the following. First, u_{1j} followed a univariate normal distribution because level 2 of the PNRCT model did not have another error term. Second, the error distribution at level 1 was not necessarily the same as the error distribution at level 2. Error distribution conditional on the model could be normal at

level 1 (student level), but at level 2 it could be a heavy-tailed distribution (e.g., the existence of a large group of schools whose students perform very well and the existence of a large group of schools whose students perform very badly may create a heavy-tailed level 2 error distribution conditional on the model). In addition, one of the model assumptions states that error distributions at level 1 and at level 2 are not correlated. Thus, the error distribution at level 2 did not have to be the same as that at level 1. Third, the resulting Y_{ij} distribution was not the focus of the study; the focus was the conditional u_{ij} distribution. The Y_{ij} values were generated using equations (1.8) and (2.29):

$$Y_{ij} = \gamma_{00} + \gamma_{10}T_{ij} + u_{1j}T_{ij} + r_{ij}$$

and

$$Y_{ij} = \gamma_{00} + \gamma_{10}T_{ij} + \gamma_{20}X_{ij} + u_{1j}T_{ij} + r_{ij}, \quad \text{respectively.}$$

Data for this study were then simulated using R software version 3.2.2 (R core Team, 2015) following the next steps for each cell in the factorial design:

Step 1. Generating values of level 1 predictors. First, T_{ij} and X_{ij} values were simulated. An equal number of zero (control) and one (treatment) values for T_{ij} were created to denote the treatment and control conditions. The values for each group were the cluster sizes (6, 17 and 32) selected in section 3.4.1. These values were then assigned to the treatment and control groups. The X_{ij} values were generated (for Model B) from a standard normal distribution ($X_{ij} \sim N(0,1)$) and were randomly assigned to T_{ij} values. Second, clusters were randomly generated in the treatment group.

Step 2. Generating level 1 error terms (r_{ij}). The next step was to simulate the level 1 error term distribution following a normal distribution with mean zero and

constant variance, which is $r_{ij} \sim N(0, \sigma^2)$. Note that σ^2 took different values in different cells, because the variance was a function of τ_{11} and the ICC. In addition, note that the X_{ij}, r_{ij}, and T_{ij} were independent as a result of the way data were simulated. This is consistent with the assumptions of HLM.

Step 3. Generating level 2 error terms (u_{1j}). This step consisted of simulating the level 2 error terms (u_{1j}) from a known distribution with mean zero and variance τ_{11}. Each error term was then randomly assigned to each cluster in the treatment condition. Remember that τ_{11} was 1 only for the normal distribution; for the t distribution with four degrees of freedom and the t distribution with 11 degrees of freedom the value of τ_{11} was determined by the number of degrees of freedom, as previously mentioned. Thus, τ_{11} was 1.22 for the t distribution with 11 degrees of freedom and 2 for the t distribution with 4 degrees of freedom.

It is worth noting that in Model A and Model B, u_{0j} and u_{2j} did not produce any variance (τ_{00} or τ_{22}) or covariance terms (τ_{01}, τ_{02}, and τ_{12}). This was because β_{0j} (in Model A) and β_{2j} (in Model B) were fixed but β_{1j} was random. As a result of this, in equations (1.8) and (2.29), the respective covariance of u_{0j} and u_{2j} in the variance-covariance matrices was zero.

Step 4. Generating level 1 outcome variable (Y_{ij}). Finally, the pre-defined values of the fixed effects and the simulated values for T_{ij}, X_{ij}, u_{1j}, and r_{ij} were substituted into equations (1.8) and (2.29) to produce the values of the outcome variable (Y_{ij}) for Model A and Model B, respectively. Notice that when producing the outcome variable for Model A, X_{ij} values were not substituted in equation (1.8). This substitution was only used when producing Y_{ij} for Model B.

Step 5. Replication process. Steps 1 to 4 were repeated for each model and each replication within a cell, and this data-generation procedures replicated to the next cell and continued until all cells were completed.

Although specific software for fitting the PNRCT model or a generalized PNRCT model is unavailable, they can still be modeled using software for fitting linear mixed models. Within R, the lme4 package (Bates et al., 2015) was used to fit the PNRCT models, and the lmerTest package (Kuznetsova, et al., 2015) was used to generate the respective *p*-values of each test. Parameter estimates, their standard errors, degrees of freedom, t-values, p-values, level 1 variance, and level 2 variance for each cell were extracted and saved in an external text file for the subsequent analysis. To facilitate further analysis and validation, all simulated datasets were also saved in external text files.

3.7. Number of Replications

There is not a specific protocol to follow when determining the number of replications. Hauck and Anderson (1984) reviewed a large number of simulation studies and concluded that researchers frequently do not follow any specific procedure for this issue. In the HLM and PNRCT fields, the same situation is usually found. For instance, some studies do not offer any explanation about how researchers defined the number of replications (Candel & Van Braukelen, 2009; Ketelsen, 2014; Lou et al., 2014; Sander, 2011; Tesller, 2014), but other studies such as Baldwin et al. (2011) and Korendijk (2012) selected the number of replications for minimizing the standard errors and for prioritizing power, respectively. As a result of this issue, a different number of

replications can be found across simulation studies. For instance, some simulation studies investigating hierarchical linear models with two levels have used between 500 and 3000 replications per cell (e.g., Ketelsen, 2014; LeBeau, 2013; Maeda, 2007; Zhang, 2005). On the other hand, simulation studies that focus on PNRCT models have used between 1000 and 10000 replications (e.g., Baldwin et al., 2011).

In this research, the number of 1000 replications was selected for several reasons. First, this number has been used frequently in PNRCT simulation studies. Second, using a larger number of replications (e. g., 5000) would require much more time to complete a run. Finally, 1000 should be large enough to provide stable estimates of outcomes because most of them are means (e.g., average RB, RMSE, Type I error, Power).

3.8. Methods of Estimation

Researchers strongly recommend the estimation of random effects by using maximum likelihood (Full or Restricted), which relies on the assumption of normality and large sample theory (Delphis, 2006; Mass & Hox, 2004b). This may be the reason why maximum likelihood is the most frequently used method of estimation

The PNRCT model requires estimating fixed effects (γs) and random effects (σ^2, and τ_{11}). For the fixed parameters, the estimation is performed by one of several methods depending on the software. For instance, HLM (Raudenbush, Bryk, & Congdon, 2011) estimates fixed parameters by using generalized weighted least squares, but lme4 (Bates, Maechler, Bolker, & Walker, 2015) uses the Maximum Likelihood (ML) estimation. When estimating the random effects, both HLM and lme4 use two ML estimation methods: Full Maximum Likelihood (FML) and Restricted Maximum Likelihood (REML) (Hox, 2002; Raudenbush & Bryk, 2002; West, Welch, Gatecki, 2007).

Researchers report several methods for implementing FML or REML, but the most common are the expectation-maximization (EM) algorithm, the Newton-Raphson (N-R) algorithm, and the Fisher Scoring algorithm; lmer function (from lme4 package) uses, depending on the selection, EM and N-R algorithms (West, Welch, & Gatecki, 2007).

In general, the ML function estimates unknown parameters by optimizing a given function. So the first step in using FML is to construct the likelihood function based on the model parameters specified in the model and the distributional assumptions. REML does the same as FML, and REML produces similar results when there is a large number of level 2 units. Despite the fact that FML and REML produce similar results for the random effects at level 1, results differ when estimating variance components at level 2 (Raudenbush & Bryk, 2002). This happens because REML adjusts the estimation for the degrees of freedom, producing better estimates than FML. Thus, REML is more efficient than FML because REML produces unbiased variance components (Harville, 1977; Browne, 1998).

In most PNRCT model simulation studies, the REML method is used. However, the study performed by Candel and Van Breukelen (2009) used both ML and REML, because both methods were part of the specified conditions. Table 5 summarizes the estimation methods in PNRCTM.

Table 5

Summary of Methods of Estimation in PNRCT Simulation Studies

Study	Estimation Method
Baldwin et al. (2011)	REML
Korendijk, et al. (2012)	REML
Tesller (2014)	REML
Sander (2011) First study	ML
Sander (2011) Second study	ML
Luo, Cappaert and Ning (2015)	REML
Candel and Van Breukelen (2009)	ML and REML

In the present study REML was used. REML estimation may perform well given the characteristics of the present study (relatively few clusters, 10 and 30). Therefore, the performance of the parameter estimates was evaluated when the PNRCT model was fitted with a relatively small number of clusters.

3.9. Data Analysis

The data analysis of simulation outcomes focused on the *RB* and variability (*RMSE*) of the fixed effect estimates, along with Type I error rate and power of tests of these effects. These statistics were examined visually and reported descriptively for each design factor as well as for the average within cells.

The results were also examined by performing an inferential analysis. Thus, several analyses of variance (ANOVA) were used to assess the effects of the design factors (cluster sizes, level 2 units, ICC, and error distributions) on the bias parameter estimates, Type I error rate and power. Separate ANOVAs were fitted to each outcome.

The ANOVA model is represented by the following equation:

$$\delta_{klmn} = \mu + C_{(k)} + D_{(l)} + E_{(m)} + F_{(n)} + CD_{(kl)} + CD_{(km)} + CF_{(kn)} + DE_{(lm)} + DF_{(ln)} + EF_{(mn)} + CDE_{(klm)} + CDF_{(kln)} + CEF_{(kmn)} + DEF_{(lmn)} + e_{klmn}, \quad (3.11)$$

where δ_{klmn} represents the mean of the dependent variable (e.g., RB) for the *klmn-th* cell, μ captures the grand mean, *C*, *D*, *E*, and *F* represent each design condition, and *k*, *l*, *m* and *n* represent each level for each design condition. In this model, the highest level interaction was pooled into the error term because there was only one observation per cell.

Due to the large sample size and power, a partial eta-squared was computed in order to quantify the magnitude of the effect size for all main effects and interactions. η_p^2 was computed by using the total sum of squares of the significant effect (given by the F test) in the ANOVA:

$$\eta_p^2 = \frac{SS_{strt}}{SS_{stotal}}. \quad (3.12)$$

In this equation η_p^2 was the partial eta-squared. SS_{strt} was the sum of squares of the significant effect, and SS_{stotal} was the total sum of the square of all significant effects. This effect size was interpreted following Gamst, Meyer, and Guarino's (2008) recommendations, which uses .09 as the cutoff for a small effect, .14 for a medium effect, and .22 for a large effect. Thus values below .09 were treated as negligible effects.

When plotting interaction effects, I followed Harwell's (1998) suggestion. Harwell indicates that it is necessary to adjust the mean cells to accurately characterize the nature of an interaction. This was done by fitting a regression model without including the interaction to be plotted. Then the cell mean residuals were plotted because

the cell means have had all other effects in the model removed, so any pattern will represent the interaction plus sampling error.

RMSE was transformed to ln(RMSE) and used as an outcome in a Weighted Least Square (WLS) Regression. The weights were the inverse of the $\ln(RMSE_j)$ variance. This is 1/ (2/n-Q-1), where n is the number of replications and Q is the number of predictors. This procedure was performed for the γ_{10} RMSE in Experiment 1 and γ_{10} RMSE and γ_{20} RMSE in Experiment 2. The WLS Regression model is

$$\ln(RMSE) = \beta_0 + \beta_1 C + \beta_2 D + \beta_3 E + \beta_4 F + \beta_5 CD + \beta_6 CE + \beta_7 CF + \beta_8 DE + \beta_9 DF + \beta_{10} EF + \beta_{11} CDE + \beta_{12} CEF + \beta_{13} CDF + \beta_{14} DEF + e_i, \quad (3.13)$$

here, β_i represents the effect associated with each variable, *C* is a set of two dummy variables representing the distribution type, *D* is the number of clusters, *E* the size of the clusters, and *F* is the ICC level. The remaining terms represent the interactions between variables (e.g., *CD* represents the interaction between the distribution type and number of clusters).

3.10. Summary

The present Monte Carlo study was designed to examine the performance of the PNRCT model when the assumption of normality did not hold in the level 2 error distribution. To achieve the goal of the study a 3 × 3 × 3 × 3 factorial design (81 cells) was used, including the following conditions: (a) number of clusters (10, 30 and 50); (b) cluster size (6, 17, 32); (c) ICC (.05, .15, .25); and (d) three distributions (normal, t distribution with four degrees of freedom, and t distribution with 11 degrees of freedom). The model's performance was analyzed by evaluating the relative bias, and root mean

square of error of the fixed effects (γ_{10}, γ_{20}). Additionally, a 95% confidence interval was created for assessing the Type I error rate. The study also proposed to examine the power, comparing it with a .80 cutoff.

Chapter 4

Results

In this chapter, results are reported from the analysis performed for the simulation conditions. To evaluate the model performance of the Model A and the Model B, the analyses was conducted on the relative bias of the fixed effects parameter estimates and their accuracy. Additionally, power and the Type I error rate were assessed.

One important thing to take into account for understanding the results of the present study is that the PNRCT model differ from HLM

This chapter includes four sections. The first section presents evidence that evaluates the generated data. Section two lists the results of Experiment 1 (evaluation of Model A), which deals with the pure PNRCT model. Section three discusses the results of Experiment 2 (evaluation of Model B), which deals with the PNRCT model adjusted by a covariate. For the convenience of readers, sections two and three follow the same order of presentation, results by levels of each condition and results by cells. The results by cells present (a) relative bias, (b) RMSE, (c) power, and (d) Type I error rate, all of them for the fixed effects.

4.1. Evaluation of the Generated Data Set

One important issue in Monte Carlo studies is the adequacy of the simulated data. Harwell and Kohli (2015) argued for the importance of providing evidence showing that the simulated values have the specified properties. If the generated values do not have the specified properties, the final analysis may not be valid and could lead to wrong

conclusions. For instance, if one of the conditions has a non-normal error distribution at level 1 of a PNRCT model, e.g. a t distribution, and the generated data show, or are sufficiently close to showing, t distribution properties, the analysis of the results will be valid for the study, leading to correct conclusions. Empirical evidence such as descriptive statistics and plots are presented in this subsection. Evidence for Experiment 1 of the study is presented first, followed by evidence for Experiment 2.

The tables and figures of this section provide evidence for the adequacy of the simulated data. They summarize evidence for the cell with 50 clusters, 32 observations within clusters (total sample of 1,600 observations in the treatment condition), and an ICC of .25. This condition is reported in detail because it is illustrative of other conditions in the simulation.

The data were generated based on the model presented in equation (1.8) for Experiment 1 and based on the model presented in equation (2.29) for Experiment 2. Descriptive statistics are presented for level 1 normal error distribution and for level 2 normal error distribution, t distribution with four degrees of freedom, and t distribution with eleven degrees of freedom. In addition, more evidence showing the adequacy of the simulated data is presented. First, the theoretical and observed percentage (probability) of values below and above -3 and 3 Z- and t-values of the level 2 error distributions were compared across distributions. Second, the tail weight index in equation (3.5) was estimated and compared for the observed distributions. It was expected that the tail weight of $u_{1j} \sim N(0,1)$ would be lower than the tail weight of $u_{1j} \sim t_{11df}$ (0, 1.106), which should be lower than the tail weight of $u_{1j} \sim t_{4df}$ (0, 1.414).

Table 6 and Figure 2 provide evidence for the adequacy of the simulated data in Experiment 1. The values of the mean, standard deviation, and variance for both the level 1 and 2 error distributions present negligible difference when compared with the specified values. Additionally, the value of the skewness for each distribution at each level is very close to zero, which suggests that the three level 2 error distributions are symmetric.

Table 6

Descriptive Statistics of the Observed Level 1 and 2 Error Distribution in Experiment 1

n_j = 50, m = 32, ICC = .25	Mean	SD	Variance	Skewness
		Normal distribution		
r_{ij} ~ N(0,1.732)	-.002	1.732	2.999	-0.002
u_{1j} ~ N(0,1)	.004	1.005	1.011	-0.019
		t distribution 4 df		
r_{ij} ~ N(0, 2.449)	.001	2.449	5.998	0.001
u_{1j} ~ t_{4df} (0, 1.414)	-.005	1.432	2.050	0.007
		t distribution 11 df		
r_{ij} ~ N(0, 1.914)	-.001	1.914	3.665	0.001
u_{1j} ~ tt_{11df} (0, 1.106)	.009	1.108	1.229	0.029

Note: n_j is the number of clusters, m is the cluster size, ICC is the intra-class correlation, r_{ij} and u_{1j} represent level 1 and 2 error distribution.

Similarly, the density plots in Figure 2 suggest that the observed distributions and the theoretical distributions are very alike, although the observed distributions deviate slightly from the theoretical distributions. However, this does not provide enough evidence to reject the idea that both observed and theoretical distributions are the same.

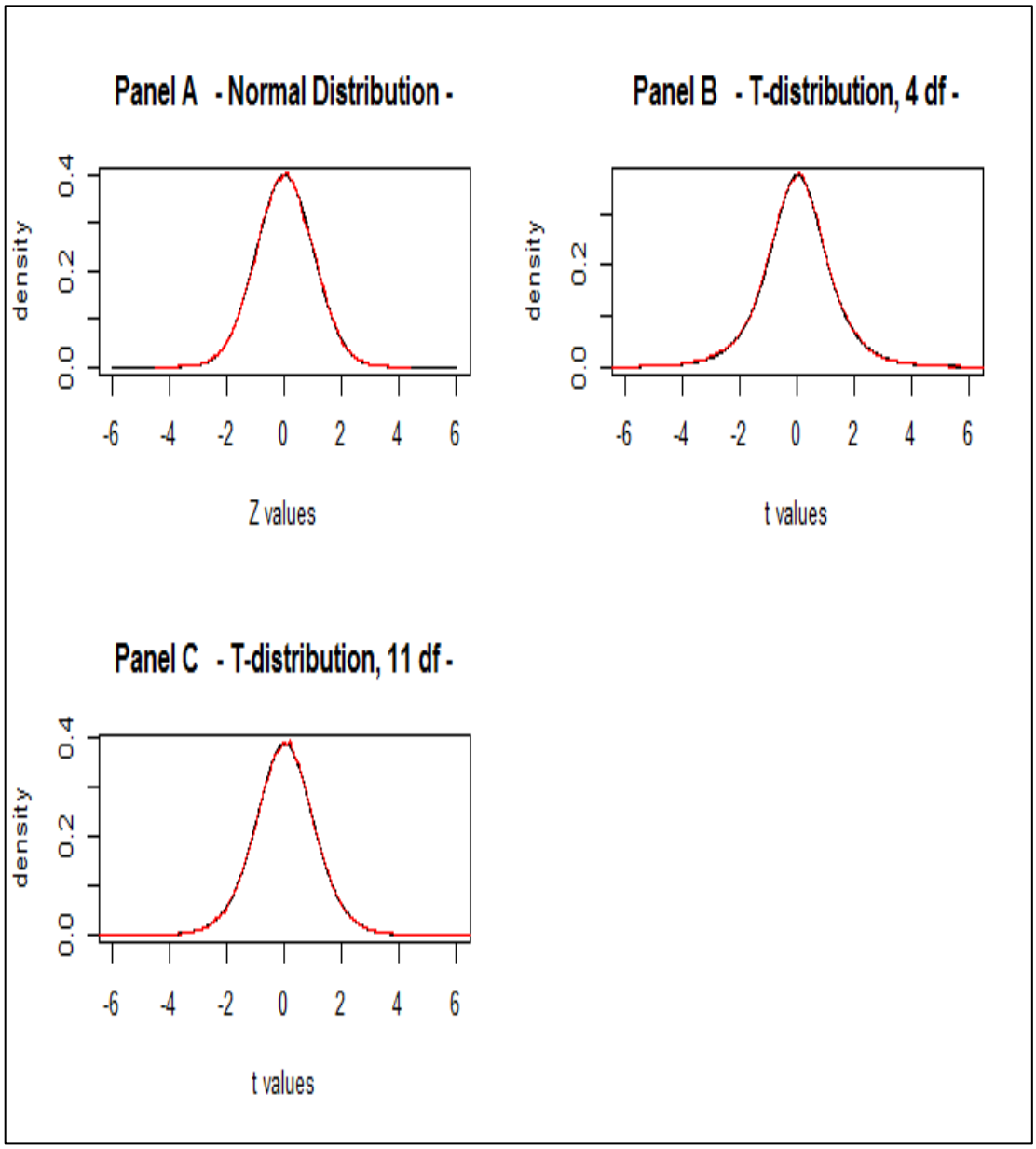

Figure 2. Density function for the observed and theoretical distributions in Experiment 1. Panel (a) sets out a normal distribution with $\bar{X}$= 0 and sd =1. Panel (b) sets out a t distribution with four degrees of freedom, whereas panel (c) shows a t distribution with eleven degrees of freedom. Red lines denote the observed distributions and black lines the theoretical distributions.

The percentage of values for the observed and the theoretical distributions are very similar. The observed value below -3 and above 3 Z-values for the normal (t-values

for the t distributions) error distribution is .0029 whereas for the theoretical distribution it is .0027. Additionally, the observed t distribution with four degrees of freedom presents a value of .0402; meanwhile, for the theoretical distribution the value is .0399. Finally, the observed t distribution with eleven degrees of freedom presents a value of .0125 and the respective theoretical distribution presents a value of .0121. When comparing the tail weights, the results are as expected. The tail weight of the normal distribution (2.592) is lower than the tail weight of the t distribution with eleven degrees of freedom (2.776), which is lower than the t distribution with four degrees of freedom (3.213).

Table 7 and Figure 3 provide evidence for the adequacy of the simulated data for Experiment 2. The values of the mean, standard deviation, and variance for both the level 1 and 2 error distributions present negligible differences when compared with the mean, standard deviation, and variance of the specified parameters. Additionally, the value of the skewness for each distribution at each level (with one exception) is close to zero, suggesting that the error distributions are symmetric. The t distribution with four degrees of freedom presents a skewness larger than the other distributions, but its value deviates slightly negatively from zero. However, this does not significantly impact the simulated data, especially because, as Figure 3 shows, the observed and theoretical distributions are quite similar.

Table 7

Descriptive Statistics of the Observed Level 1 and 2 Error Distribution in Experiment 2

	Mean	SD	Variance	Skewness
n_j = 50, m = 32, ICC = .25				
		Normal distribution		
r_{ij} ~ N(0,1.732)	-.001	1.732	2.999	0.000
u_{1j} ~ N(0,1)	.0002	1.002	1.004	0.007
		t distribution 4 df		
r_{ij} ~ N(0, 2.449)	-.001	2.448	5.994	0.000
u_{1j} ~ t_{4df} (0, 1.414)	.003	1.436	2.064	-0.430
		t distribution 11 df		
r_{ij} ~ N(0, 1.915)	-.003	1.914	3.666	0.001
u_{1j} ~ t_{11df} (0, 1.106)	.002	1.108	1.228	0.001

Note: n_j is the number of clusters, m is the cluster size, ICC is the intra-class correlation, r_{ij} and u_{1j} represent level 1 and 2 error distribution.

Figure 3 shows three density plots for the observed and theoretical distributions of level 2 error distribution. It can be seen that the observed distributions deviate very slightly from the theoretical distributions. However, both observed and theoretical distributions seem to be very similar. Therefore, there is no reason to reject the idea that both observed and theoretical distributions are quite similar.

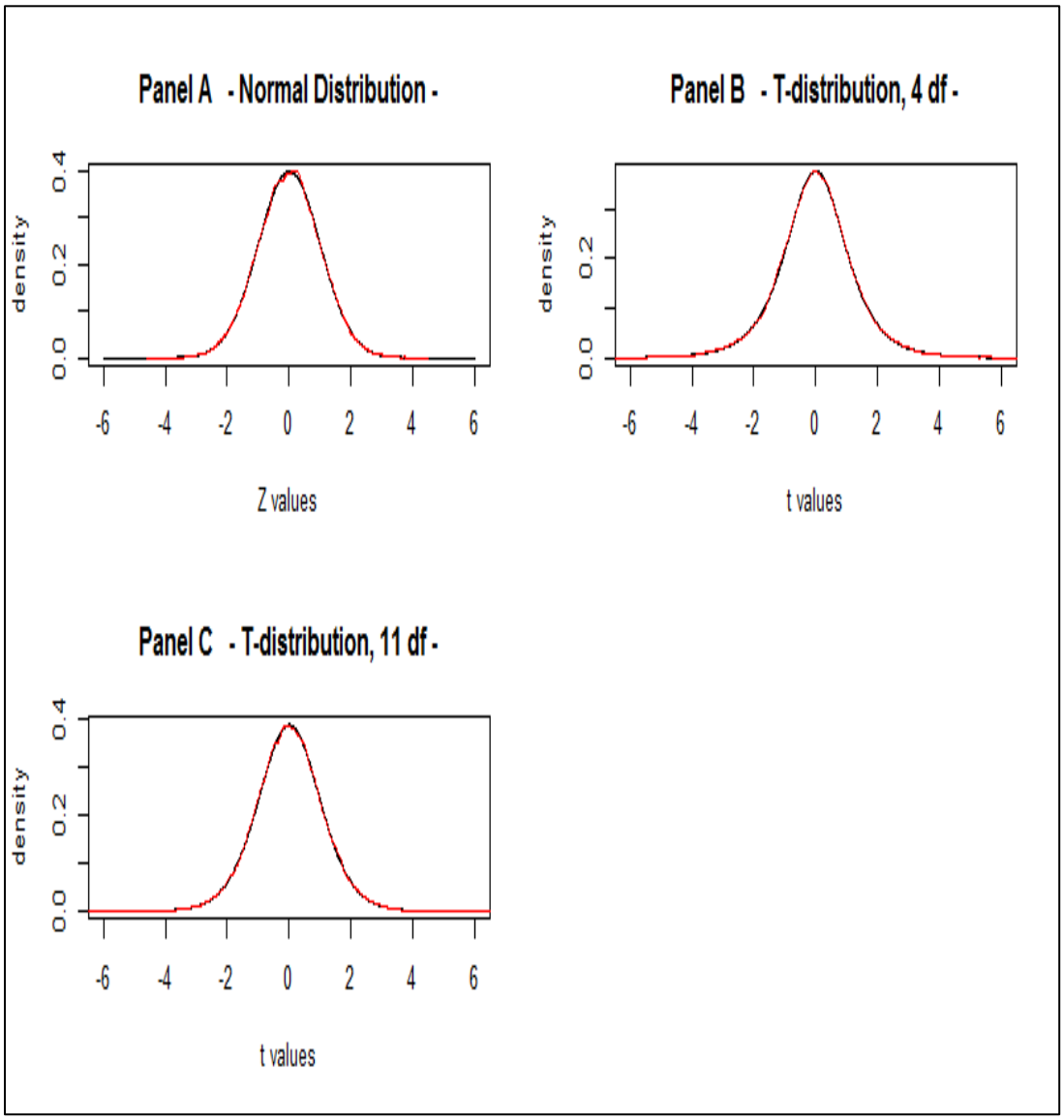

Figure 3. Density function for the observed and theoretical distributions in Experiment 2. Panel (a) shows a normal distribution with $\bar{X}$= 0 and sd =1. Panel (b) shows a t distribution with four degrees of freedom, whereas panel (c) shows a t distribution with eleven degrees of freedom. Red lines denote the observed distributions and black lines the theoretical distributions.

The observed and theoretical percentages (probabilities) below -3 and above 3, for the normal error distribution are the same. The observed distribution has a value of .0027,

whereas the theoretical distribution also has a value of .0027. However, the observed t distribution with four degrees of freedom has a value of .0402, while the theoretical distribution has a value of 0.040. In addition, the observed t distribution with eleven degrees of freedom has a value of .012, and the respective theoretical distribution has a value of .012. Finally, the tail weight of the normal distribution (2.586) is lower than the tail weight of the t distribution with eleven degrees of freedom (2.775), which is lower than the t distribution with four degrees of freedom (3.237).

The above results confirm what the literature indicates: a t distribution has heavier tails than a normal distribution (Daniel, 2005). The percentages (probabilities) of the observed normal error distribution below and above -3 and 3 Z-value (t-values for the t distributions) are lower than the values of error for a t distribution with four and eleven degrees of freedom. In addition, both t distributions with four and eleven degrees of freedom show Q indices larger than the normal distribution, suggesting that they have heavier tails than the normal distribution.

Although all evidence presented above does not perfectly match with the theoretical parameters, the estimated parameters in Tables 6 and 7 and the plots in Figure 2 and 3 suggest that the data for this study was simulated adequately.

4.1.1. Accuracy of fixed effects. As previously mentioned, the data were generated following normal error distribution at level 1 and normal error distribution and t distributions with four and eleven degrees of freedom at level 2. If the data values were properly generated, the average parameter estimates should have similar values as the corresponding specified parameter values over replications. Thus to illustrate the

accuracy of the parameter estimates, these parameter estimates were averaged in the cell with 50 clusters, 32 observations (clusters size), and ICC of .25. Table 8 lists the descriptive statistics of the parameter estimates for each distribution.

Table 8

Descriptive Statistics of the Fixed Effects by Distribution in Experiment 1

Distribution	Parameter estimate	Mean	Min	Max	SD	Skewness
Normal distribution	γ_{10}	1.009	0.356	1.423	0.157	-0.194
t_{11df} distribution	γ_{10}	1.008	0.448	1.466	0.167	-0.100
t_{4df} distribution	γ_{10}	0.991	0.332	1.703	0.217	-0.069

Note: Parameter statistics were estimated over 1,000 data sets. Min stands for minimum value, Max for maximum value, and SD for standard deviation. The specified value of γ_{10} was 1.

The estimation of the descriptive statistics was performed by averaging 1,000 parameter estimates for each distribution. The fixed effects seem to be fairly estimated. Their values are very close to the specified value, $\gamma_{10} = 1$. Furthermore, the distribution of the parameter estimates ranges between 0.332 and 1.703, but overall the distribution seems to be somewhat symmetric (See Figure 4). The standard deviation shows that the distribution of the fixed effects for each distribution are relatively narrow. In addition, skewness has negligible values for each distribution of the fixed effects. Figure 4 illustrates this fact for the results of γ_{10}.

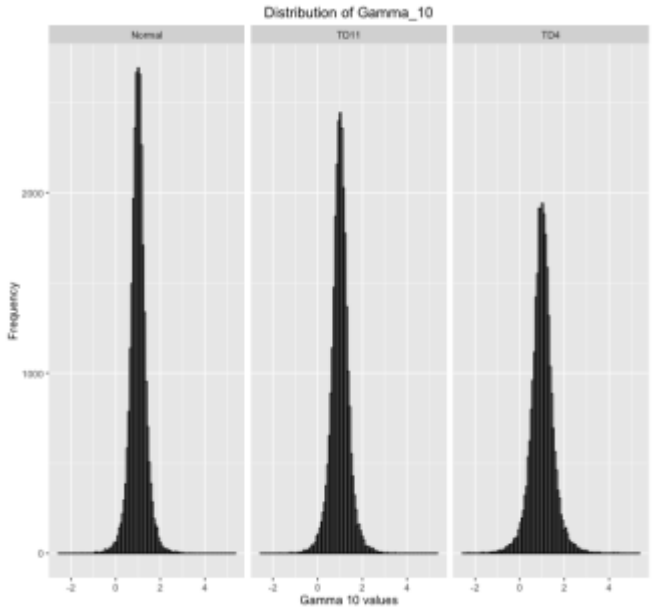

Figure 4. Observed distribution of γ_{10} in Experiment 1. Panel A shows the γ_{10} generated under normal distribution. Panel B shows the γ_{10} generated under t distribution with 11 degrees of freedom. Panel C shows the γ_{10} generated under t distribution with 4 degrees of freedom.

The same procedure was followed for Experiment 2 to illustrate the accuracy of the parameter estimates. The results are presented in Table 9, which shows the descriptive statistics by distribution for γ_{10} and γ_{20}.

Table 9

Descriptive Statistics of the Fixed Effects by Distribution in Experiment 2

Distribution	Parameter estimate	Mean	Min	Max	SD	Skewness
Normal distribution	γ_{10}	0.996	0.563	1.546	0.154	0.090
	γ_{20}	1.000	0.897	1.095	0.031	0.019
t_{11df} distribution	γ_{10}	1.001	0.428	1.564	0.171	-0.038
	γ_{20}	0.999	0.881	1.124	0.034	0.009
t_{4df} distribution	γ_{10}	0.997	0.283	1.757	0.219	0.154
	γ_{20}	1.002	0.859	1.145	0.045	0.055

Note: parameter statistics were estimated over 1000 data sets. Min stands for minimum value, Max for maximum value, and SD for standard deviation. The specified values for both γ_{10} and γ_{20} was 1.

The estimation of the descriptive statistics in Experiment 2 was performed by averaging 1,000 parameter estimates for each distribution in the cell with 50 clusters, 32 observations within clusters, and ICC of .25, as for Experiment 1. The results show that the fixed effects values are very close to the specified values $\gamma_{10} = 1$ and $\gamma_{20} = 1$. Furthermore, the distributions of γ_{10} and γ_{20} present different ranges and they seem to be symmetric although they present negligible skewness. The standard deviations of the parameter estimates do not reflect a large variation. Figures 5 and 6 show the shape of the distribution of γ_{10} and γ_{20}.

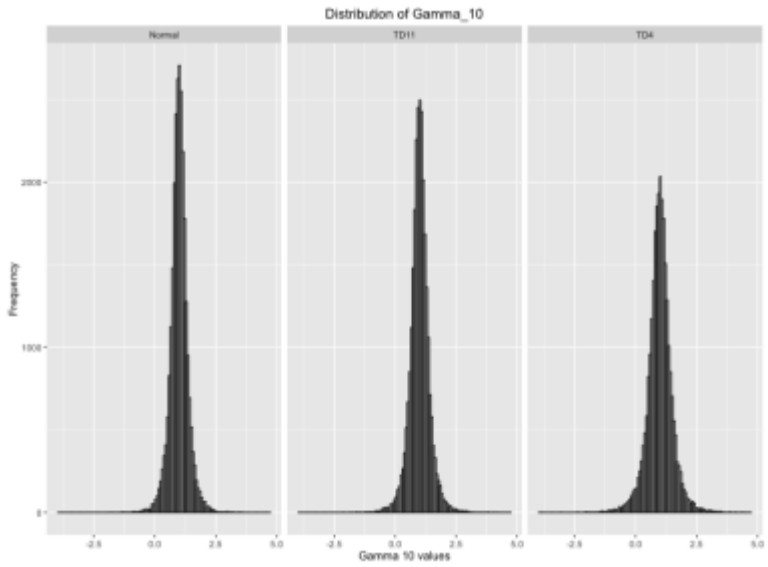

Figure 5. Observed distribution of γ_{10} in Experiment 2. Panel A shows the γ_{10} generated under normal distribution. Panel B shows the γ_{10} generated under t distribution with 11 degrees of freedom. Panel C shows the γ_{10} generated under t distribution with 4 degrees of freedom.

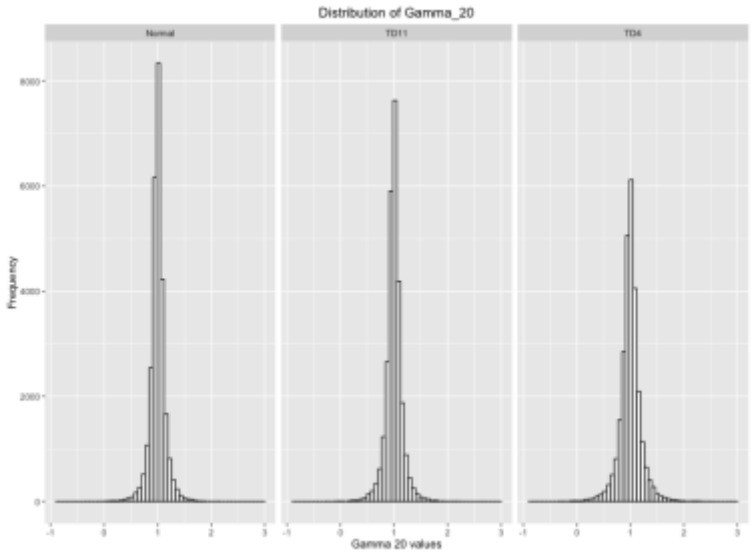

Figure 6. Observed distribution of γ_{20} in Experiment 2. Panel A shows the γ_{20} generated under normal distribution. Panel B shows the γ_{20} generated under t distribution with eleven degrees of freedom. Panel C shows the γ_{20} generated under t distribution with four degrees of freedom.

4.2. Results of Experiment 1

4.2.1. Results by levels of each condition. This section presents the results for the levels of each condition in Experiment 1, which are summarized in Table 10. The γ_{10} relative bias for each condition is very small, suggesting that γ_{10} is not biased across conditions. Additionally, RMSE is smaller in the normal distribution and t distribution with eleven degrees of freedom, and higher in the t distribution with four degrees of freedom. This suggests that parameter estimates in the t distribution with four degrees of freedom are less accurate. Examining the number of clusters shows that the RMSE value decreases as the number of clusters increases. A similar pattern is found in the cluster size and ICC conditions. These findings suggest that as the number of clusters, cluster size, and ICC increase the accuracy of the parameter estimates increases.

The results in Table 10 suggest that overall power is lower in the two t distributions (power less than .80) than the normal distribution (power approximately equal to .80). The difference is .15 between the normal distribution and the t distribution with four degrees of freedom, and 0.04 between the normal distribution and the t distribution with eleven degrees of freedom.

Additionally, when the number of clusters is ten, the power is low (less than .50) compared to that for 30 and 50 clusters (greater than .80). A similar situation occurs with power when examining the cluster size; power is low when the cluster sizes are 6 and 17 (below .80) compared with 32 (above .80). However, the difference in power among the cluster sizes is relatively low. Similarly, power is relatively low when the ICC values are .05 and .15 (below .80) compared to .25 (slightly above .80). As was explained in Chapter 3, power was expected to increase as ICC increases. This is because of the way data were simulated.

The Type I error rate does not represent a problem in any of the conditions. This statistic is in between the interval set in Chapter 3 (.36, .63).

Table 10

Marginal Effects for the Specified Conditions and Dependent Variables

		RB of γ_{10}	RMSE of γ_{10}	Power of γ_{10}	Type I Error Rate of γ_{10}
Distribution	Normal	0.001	0.361	.804	.049
	t_{11df}	-0.001	0.402	.760*	.052
	t_{4df}	0.003	0.519	.651*	.051
Number of Clusters	10	0.003	0.367	.457*	.051
	30	-0.002	0.121	.831	.050
	50	0.001	0.073	.928	.050
Cluster size	6	0.002	0.367	.618*	.049
	17	0.003	0.121	.774*	.054
	32	-0.003	0.073	.824	.049
ICC	.05	-0.001	0.367	.599*	.050
	.15	0.002	0.121	.783*	.049
	.25	0.002	0.073	.833	.053

Note: RB stands for average relative bias and RMSE for Root Mean Square Standard Error, * denotes a power lower than .80 and relative bias greater than 0.05.

4.2.2. Results by cells.

4.2.2.1. Relative bias of γ_{10}. In the previous section, the results were presented for the levels of each condition. In this section, the results are presented for the 81 cells of the study.

Table 11 summarizes the relative bias for γ_{10}. The average cell bias across all cells ranges from -0.034 to 0.043, which implies that γ_{10} deviates between ± 5% from the specified parameter value. This suggests that the relative bias of γ_{10} does not represent a problem under any condition in any cell. In addition, the results across cells do not show (visually examined) any specific pattern across conditions. Values in Table 11 have to be interpreted as the average difference between the true parameter and the observed estimated fixed effect under the conditions of the study. For instance, the value 0.023 means that on average the observed estimated fixed effect under the study conditions (t

distribution with 4 degrees of freedom, ICC = 0.15, 10 clusters, and cluster size of 7) differs from the true parameter by 2.3%.

Since the relative bias of γ_{10} ranges from -0.034 to 0.043, some variation may exist. To make an inferential analysis and to examine whether this variation is significant, an ANOVA model was performed. The inferential analysis showed that no term of the ANOVA model was statistically significant, which suggests that the variation across cells is due to randomness. This implies that none of the main conditions of the study, distributions, number of clusters, cluster size, or ICC showed any pattern of bias on the fixed effects (γ_{10}). Moreover, the inferential analysis showed no effect for any interactions of the main conditions on the bias of γ_{10} in the experiment. Table 28 in the Appendix shows the ANOVA results.

Table 11

γ_{10} Relative Bias for All Conditions in Experiment 1

	10 Clusters			30 Clusters			50 Clusters		
	Cluster size			Cluster size			Cluster size		
ICC	6	17	32	6	17	32	6	17	32
				Normal distribution					
.05	0.002	0.029	-0.034	-0.008	-0.002	-0.003	0.007	-0.001	-0.003
.15	-0.022	0.017	0.006	-0.002	0.002	0.004	0.016	-0.010	-0.005
.25	0.006	0.009	0.007	0.011	0.011	-0.012	-0.001	0.004	0.009
				t distribution 4 df					
.05	0.001	0.043	-0.005	0.008	-0.017	-0.009	0.000	0.008	-0.003
.15	0.010	0.023	0.003	0.003	-0.002	-0.020	0.003	0.011	0.001
.25	0.003	0.010	0.036	-0.005	-0.008	0.001	0.002	-0.014	-0.009
				t distribution 11 df					
.05	0.021	-0.012	-0.032	-0.016	0.010	-0.006	0.003	-0.002	-0.009
.15	0.022	-0.020	-0.005	0.016	-0.003	-0.007	-0.004	0.015	0.005
.25	-0.020	-0.007	-0.002	0.002	-0.003	0.012	0.001	-0.008	0.008

4.2.2.2. RMSE of γ_{10}. As defined in Chapter 3, the accuracy of γ_{10} was investigated by estimating the root mean square error (RMSE). Across cells the minimum and the maximum values were 0.127 and 0.961, respectively. (See Table 12). This implies that RMSE may not be uniform across conditions, suggesting that in some cells γ_{10} is estimated more accurately than in others. In fact, a visual examination suggests several patterns. First, RMSE decreases as the ICC, cluster size, and number of clusters increase. This implies that γ_{10} becomes more accurately estimated as the ICC, cluster size, and the number of clusters increase. Second, when comparing the RMSE across distributions, the values in the normal distribution are smaller than in the t distribution with four and t distribution with eleven degrees of freedom, suggesting that γ_{10} is more accurately estimated under the normal distribution. However, in some cases the differences are slight, especially between the normal distribution and the t distribution with eleven degrees of freedom.

Table 12

RMSE of γ_{10} in Experiment 1

	10 Clusters			30 Clusters			50 Clusters		
	Cluster size			Cluster size			Cluster size		
ICC	6	17	32	6	17	32	6	17	32
				Normal distribution					
.05	0.858	0.562	0.459	0.483	0.331	0.270	0.371	0.266	0.210
.15	0.510	0.416	0.383	0.310	0.228	0.209	0.243	0.184	0.165
.25	0.438	0.374	0.339	0.247	0.213	0.197	0.202	0.169	0.158
				t distribution 4 df					
.05	1.233	0.826	0.646	0.691	0.468	0.394	0.527	0.367	0.297
.15	0.786	0.614	0.533	0.431	0.338	0.294	0.334	0.253	0.228
.25	0.650	0.527	0.469	0.363	0.304	0.286	0.284	0.245	0.217
				t distribution 11 df					
.05	0.923	0.631	0.515	0.556	0.374	0.301	0.422	0.285	0.239
.15	0.592	0.459	0.399	0.340	0.262	0.239	0.273	0.207	0.181
.25	0.506	0.406	0.373	0.293	0.241	0.224	0.234	0.183	0.168

Third, the RMSE values in the normal distribution are below 0.30 (arbitrarily selected) when the number of clusters is 30 and the cluster size is 6. A similar situation happens in the t distribution with eleven degrees of freedom. By contrast, in the t distribution with four degrees of freedom this situation occurs when the number of clusters is 50 and the cluster size is 6.

As mentioned earlier, the RMSE values range from 0.127 to 0.961, suggesting some variation across cells. The WLS regression model performed on ln(*RMSE*) showed a negative effect for the number of clusters (β_2 = - 0.020, p-value < .01), cluster size (β_3 = - 0.025, p-value < .01) and ICC (β_4 = - 3.64, p-value < .01). The remaining variables did not have an impact on the ln(RMSE). These results suggest that as the number of clusters, the cluster size and the ICC increase, the RMSE decreases in 0.02, 0.025 and 3.64 units, respectively (See Table 29). In other words, the fixed effect (γ_{10}) is more precisely estimated as the number of clusters, cluster size, and ICC increase. The following figure illustrates the pattern of the main effects on the RMSE.

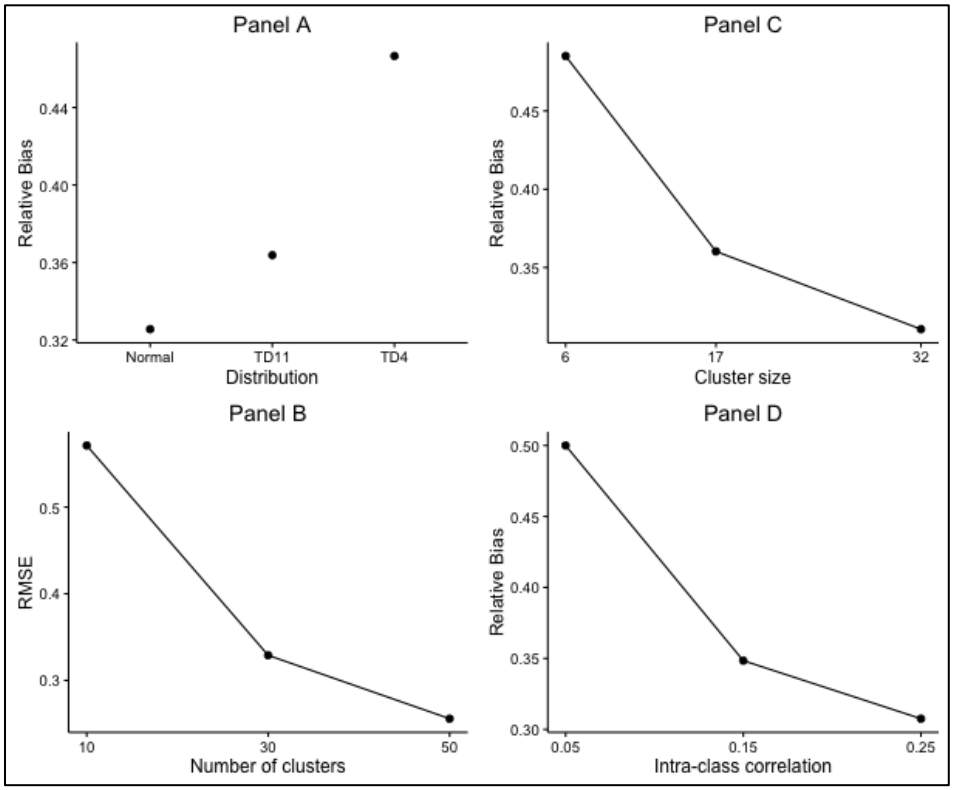

Figure 7. Main effects on γ_{10} RMSE. Panel A shows the main effect of the error distributions on RMSE. Panel B shows the main effect of the levels of number of clusters on γ_{10} RMSE. Panel C shows the main effects of the levels of cluster size on γ_{10} RMSE. Panel D shows the main effect of ICC levels on γ_{10} RSME.

4.2.2.3. Power of γ_{10}. Power by cells is presented in Table 13. These values represent, on average, the percentage of the number of times a null hypothesis is rejected, under the specified conditions, when this hypothesis is false. Take as example the interpretation of 0.768. On average, the Model A rejects the null hypothesis when it is false 76.8% of the times under a t distribution with four degrees of freedom, ICC = 0.25, and cluster size of 6 and 30 clusters.

From this Table 13, some patterns can be seen. First, power increases as the number of clusters and the cluster size increase. The combination of cluster and cluster size provides the total sample; thus as the sample size increases, power also increases.

This situation is consistent with the literature, which indicates that power is a function of the sample size, among other factors (Cohen, 1969).

Second, in the normal distribution a power of .80 (non-conservative value) is reached when the total sample size is 510 individuals, which is equivalent to 30 clusters with 17 observations within clusters. This happens when the ICC is .05. However, when the ICC is .15 or .25, a power of .80 is reached with a total sample size of 180 observations, which is equivalent to 30 clusters with cluster size of 6.

A similar pattern is found in the t distribution with eleven degrees of freedom. However, in this distribution with an ICC of .05, a power value of .80 is reached when the total sample size is 960, which is equivalent to 30 clusters with 32 observations within each cluster. In addition, when the ICC is .15 or .25, a power of .80 is reached at the same threshold as that reached by the normal distribution. In the case of the t distribution with four degrees of freedom, a power of .80 is reached at a total simple size of 1600 observations, which is equivalent to 50 clusters and 32 observations within each cluster. However, when the ICC is .15 or .25, a power of .80 is reached at a total sample size of 517, which is equivalent to 30 clusters and 17 observations within each cluster. Finally, Table 13 shows that power increases as the ICC increases. This pattern can be easily seen across all conditions.

These patterns suggest that the PNRCT model is underpowered, especially when the number of clusters is 10 and the cluster sizes are 6, 17 and 32. This also happens when the number of clusters are 30 and 50 and the cluster size is the lowest. Note that this situation is exacerbated when the level 2 error distributions are non-normal.

Table 13

Power of γ_{10} by Conditions Experiment 1

	10 Clusters			30 Clusters			50 Clusters		
	Cluster size			Cluster size			Cluster size		
ICC	6	17	32	6	17	32	6	17	32
				Normal distribution					
.05	.194*	.418*	.508*	.510*	.841	.961	.755*	.972	.995
.15	.405*	.640*	.691*	.888	.986	.996	.983	1.000	1.000
.25	.556*	.705*	.741*	.975	.996	.998	.997	1.000	1.000
				t distribution 4 df					
.05	.119*	.235*	.323*	.291*	.577*	.725*	.445*	.792*	.903
.15	.254*	.435*	.478*	.636*	.833	.884	.835	.964	.992
.25	.350*	.489*	.554*	.768*	.886	.921	.932	.980	.981
				t distribution 11 df					
.05	.169*	.326*	.428*	.426*	.772*	.899	.657*	.941	.988
.15	.374*	.522*	.634*	.825	.960	.980	.958	1.000	1.000
.25	.473*	.646*	.666*	.926	.981	.992	.991	.999	1.000

Note: * denotes a power less than .80

As was predicted in Chapter 3, as ICC increases, power also increases. Educational researchers have reported an opposite pattern, which shows that as ICC increases, power decreases (e.g., Bray & Kehle, 2011; Jason & Glenwick, 2016). The reason for this pattern between ICC and power was due to the lack of control on σ^2 across levels of ICC. As mentioned in Chapter 3, it was not possible to have control over ICC, σ^2, and τ_{11} at the same time; at least one of these three parameters could not be controlled (see equation 2.20). Therefore, in this research ICC and τ_{11} were controlled, and σ^2 was allowed to be a function of ICC and τ_{11}. Since ICC had three levels, different values of σ^2 were calculated for each level of ICC.

In addition, the t distributions with four and eleven degrees of freedom had different τ_{11} each. This was because each t distribution had different variance. It is important to remember that the variance of the t distribution is estimated by $\nu/(\nu - 2)$, where the ν stands for degrees of freedom. Thus it was impossible to have the same τ_{11} across distributions.

To support the fact that σ^2 was a function of ICC and τ_{11}, Tables 14, 15, and 16 present evidence that the simulation process produced observed σ^2, ICC, and τ_{11} similar to the specified values. The specified values of σ^2 were those presented in Table 1. Meanwhile the τ_{11} values were 1, 2, and 11/9 (1.222) for the normal and t distributions with four and eleven degrees of freedom, respectively. Finally, the ICC values were .05, .15 and .25.

Table 14

Observed Values of ICC Across Conditions

	10 Clusters			30 Clusters			50 Clusters		
	Cluster size			Cluster size			Cluster size		
ICC	6	17	32	6	17	32	6	17	32
				Normal distribution					
0.05	0.063	0.049	0.049	0.051	0.050	0.049	0.049	0.049	0.050
0.15	0.144	0.144	0.146	0.146	0.150	0.149	0.149	0.149	0.149
0.25	0.233	0.238	0.242	0.245	0.247	0.246	0.246	0.246	0.250
				t distribution 4 df					
0.05	0.061	0.051	0.046	0.053	0.050	0.049	0.050	0.050	0.050
0.15	0.137	0.132	0.135	0.140	0.143	0.144	0.140	0.140	0.150
0.25	0.215	0.221	0.218	0.235	0.238	0.236	0.240	0.240	0.240
				t distribution 11 df					
0.05	0.067	0.051	0.048	0.054	0.051	0.050	0.050	0.050	0.050
0.15	0.147	0.143	0.144	0.147	0.149	0.147	0.150	0.150	0.150
0.25	0.239	0.227	0.237	0.242	0.245	0.246	0.250	0.250	0.250

Table 15

Observed Values of σ^2 Across Conditions

	10 Clusters			30 Clusters			50 Clusters		
	Cluster size			Cluster size			Cluster size		
ICC	6	17	32	6	17	32	6	17	32
				Normal distribution					
0.05	18.919	18.866	18.980	19.012	18.988	18.992	19.003	18.976	19.004
0.15	5.635	5.643	5.657	5.675	5.672	5.676	5.660	5.664	5.662
0.25	2.990	3.002	2.989	2.999	2.995	2.999	2.999	2.998	3.000
				t distribution 4 df					
0.05	37.574	37.870	37.963	37.802	37.960	38.031	37.847	37.964	37.977
0.15	11.301	11.369	11.345	11.366	11.351	11.344	11.345	11.325	11.326
0.25	5.974	5.976	6.001	5.990	5.996	5.999	5.999	6.000	5.998
				t distribution 11 df					
0.05	23.081	23.108	23.164	23.102	23.177	23.224	23.097	23.200	23.218
0.15	6.839	6.897	6.946	6.945	6.935	6.917	6.902	6.925	6.917
0.25	3.672	3.657	3.672	3.667	3.660	3.664	3.672	3.664	3.665

Note: $\sigma^2 = 19$ for an ICC value of .05; $\sigma^2 = 5.667$ for an ICC value of .15; and $\sigma^2 = 3$ for an ICC value of .25, for the normal distribution. $\sigma^2 = 38$ for an ICC value of .05; $\sigma^2 = 11.333$ for an ICC value of .15; and $\sigma^2 = 6$ for an ICC value of .25, for the t distribution with four degrees of freedom. $\sigma^2 = 23.222$ for an ICC value of .05; $\sigma^2 = 6.926$ for an ICC value of .15; and $\sigma^2 = 3.667$ for an ICC value of .25, for the t distribution with 11 degrees of freedom.

Table 16

Observed Values of τ_{11} for the Normal Distribution

	10 Clusters			30 Clusters			50 Clusters		
	Cluster size			Cluster size			Cluster size		
ICC	6	17	32	6	17	32	6	17	32
				Normal distribution					
0.05	1.355	1.009	0.995	1.060	1.017	0.989	1.006	0.978	1.012
0.15	1.028	1.007	1.007	1.000	1.017	1.009	1.008	1.001	1.001
0.25	1.007	1.005	1.012	1.000	1.007	0.998	0.999	0.993	1.015
				t distribution 4 df					
0.05	2.667	2.184	1.951	2.224	2.034	1.996	2.153	2.009	2.010
0.15	2.097	1.913	1.956	2.051	2.002	1.985	1.956	1.979	1.998
0.25	2.032	2.049	1.922	1.988	2.042	1.983	1.978	1.947	2.058
				t distribution 11 df					
0.05	1.787	1.286	1.198	1.351	1.263	1.223	1.294	1.196	1.259
0.15	1.296	1.222	1.235	1.233	1.238	1.211	1.209	1.227	1.205
0.25	1.309	1.167	1.220	1.219	1.222	1.228	1.222	1.219	1.233

Note: $\tau_{11} = 1$ for the normal distribution. $\tau_{11} = 2$ for the t distribution with 4 degrees of freedom. $\tau_{11} = 11/9$ (1.222) for the t distribution with 11 degrees of freedom.

The inferential analysis, regarding power of γ_{10}, indicates that three conditions have statistically significant variation. These conditions are number of clusters (η_p^2=.607), the ICC (η_p^2=.150), and the cluster size (η_p^2=.113). The number of clusters then presents a large effect size, suggesting that this condition may be an important determinant of power. By contrast, cluster size and ICC present small effect sizes. The remaining conditions show negligible effect size (see Table 31 in the Appendix). To visualize how power is affected by the number of clusters, cluster size and ICC, see Figure 8.

The ICC also showed an impact on the power of the fixed effects. The pattern suggests that the power of the fixed effects increases as the ICC increases. This pattern was confirmed by the inferential analysis, which indicated statistically significant variation on the power of γ_{10} across the levels of ICC. However, the positive relationship of ICC and power is due to the way ICC was controlled in the experiments, as explained in Chapter 3. It is important to mention that the distribution type showed statistically significant variation. However, its effect size was negligible ($\eta_p^2 < .09$).

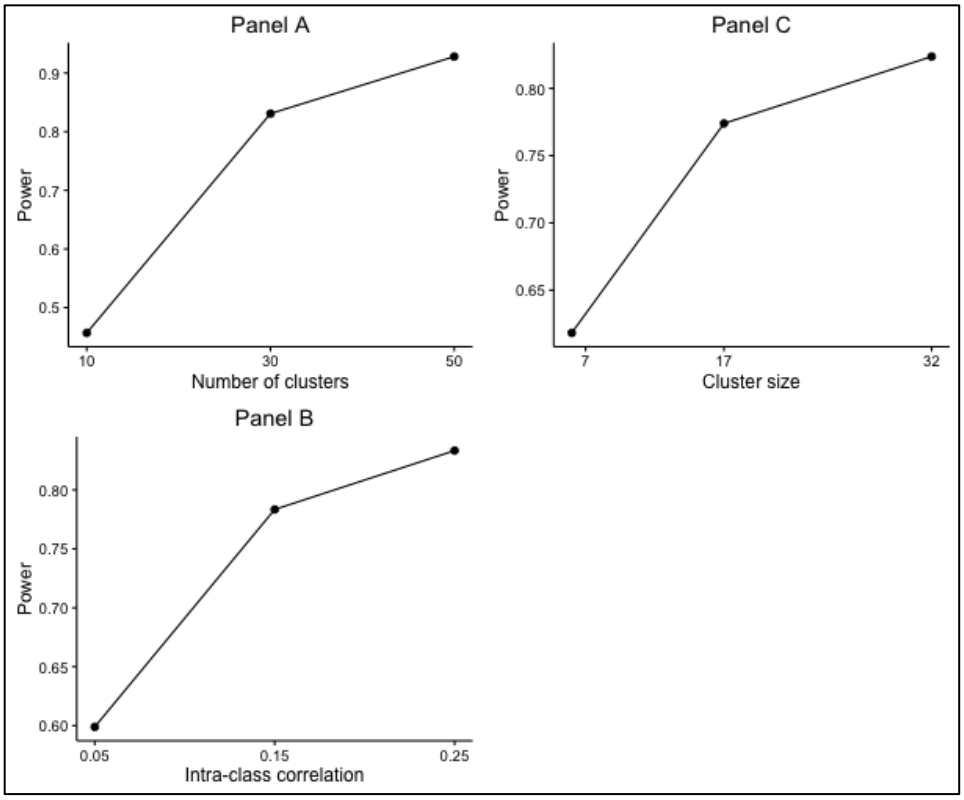

Figure 8. ANOVA conditions' main effects on γ_{10} power. Panel A shows the main effect of the levels of number of clusters on power. Panel B shows the main effect of ICC levels on power. Panel C shows the main effects of the levels of cluster size on power.

4.2.2.4. Type I error rate of γ_{10}. The Type I error rate for γ_{10} was estimated between .036 and .063 in almost all the cells of the study (see Table 17). This situation is the same in the normal distribution across all other conditions. In the case of the t distribution with four degrees of freedom, the Type I error rate is upward estimated (.063) in only one cell, whereas the t distribution with eleven degrees of freedom presents an average Type I error rate that is downward estimated (.032) in one cell, and upward estimated (.066 and .069) in two cells (see Table 17). When visually examining Table 17, no patterns are identified. This suggest that the Type I error rates outside the range of .036 to .063 are randomly downward or upward estimated.

Values in Table 17 represent the percentage of the number of times the null hypothesis is rejected when this is true. Readers could follow the next example when interpreting values in Table 17. The value 0.056 means that on average the percentage of times that the true null hypothesis is rejected is 5.6%.

To confirm no pattern in the Type I error, an inferential analysis was performed. The inferential analysis of the observed Type I error rate of γ_{10} suggests that the variation across cells is due to random error because none of the conditions showed a statistically significant variation.

Table 17

Type I Error Rate of γ_{10} by Conditions in Experiment 1

	10 Clusters			30 Clusters			50 Clusters		
	Cluster size			Cluster size			Cluster size		
ICC	6	17	32	6	17	32	6	17	32
				Normal distribution					
.05	.050	.045	0.044	.043	.047	.048	.042	.053	.048
.15	.043	.055	0.058	.051	.042	.044	.049	.053	.052
.25	.046	.055	0.054	.037	.060	.045	.049	.049	.050
				t distribution 4 df					
.05	.057	.053	0.044	.051	.055	.060	.050	.060	.047
.15	.051	.062	0.045	.043	.056	.044	.047	.057	.039
.25	.066*	.062	0.038	.045	.046	.046	.042	.060	.044
				t distribution 11 df					
.05	.032*	.053	0.053	.054	.057	.060	.045	.043	.050
.15	.046	.044	0.034	.047	.047	.056	.066*	.048	.048
.25	.059	.069*	0.056	.059	.055	.059	.061	.062	.045

Note: * indicates Type I error rate downward or upward estimated.

4.3. Results of Experiment 2

4.3.1. Results by levels of each condition. The results of Experiment 2 across the conditions levels are presented in Table 18. The relative bias values of γ_{10} are very small in the levels of each condition. These values are lower or higher by at most .05,

suggesting that γ_{10} is not biased. The relative bias of γ_{20} also presents small values, suggesting that γ_{20} is not biased. These values are even lower when compared to the relative bias of γ_{10}. Additionally, Table 18 shows that the RMSE of γ_{10} is small in the normal distribution compared with the t distribution with four and the t distribution with eleven degrees of freedom. However, the t distribution with four degrees of freedom presents the highest value among the three distributions. For the remaining conditions, number of clusters, cluster size, and ICC, the RMSE decreases as the level of each condition increases. The RMSE of γ_{20} presents a similar pattern to that presented by the RMSE of γ_{10}. However, a direct comparison of the RMSE of γ_{10} and γ_{20} shows that the RMSE of γ_{20} is smaller than the RMSE of γ_{10} under all conditions.

Table 18

Marginal Effects for the Specified Conditions and Dependent Variables

		RB γ_{10}	RB γ_{20}	RMSE γ_{10}	RMSE γ_{20}	Power γ_{10}	Type I Error γ_{10}	Power γ_{20}	Type I Error γ_{20}
Distribution	Normal	-0.003	-0.001	0.363	0.146	.799*	.051	.986	.047
	t_{11df}	0.000	-0.001	0.402	0.159	.764*	.049	.981	.052
	t_{4df}	-0.009	-0.002	0.513	0.205	.647*	.050	.959	.050
Groups	10	-0.010	-0.002	0.603	0.242	.448*	.052	.934	.051
	30	-0.001	-0.003	0.348	0.137	.834	.050	.993	.048
	50	-0.001	0.000	0.268	0.107	.929	.049	.999	.050
Within Observations	6	-0.005	-0.004	0.539	0.242	.616*	.048	.935	.050
	17	-0.004	-0.001	0.391	0.141	.771*	.051	.991	.051
	32	-0.003	0.001	0.338	0.101	.824	.051	1.000	.047
ICC	.05	-0.005	-0.004	0.551	0.245	.601*	.047	.932	.048
	.15	-0.002	-0.001	0.379	0.137	.783*	.052	.994	.051
	.25	-0.005	0.000	0.331	0.099	.828	.051	1.000	.050

Note: * denotes a power lower than .80 and relative bias greater than 0.05.

The power of γ_{10} is below .80 in the three distributions. However, in the normal distribution, power presents the highest value among the three distributions. The difference between the observed power and .80 in the normal distribution is negligible. The difference between the observed power of the t distributions with four and eleven degrees of freedom and .80 may be considered moderate and small, respectively. When examining power in the condition of number of clusters, power increases as the number of clusters increases. For ten clusters, power is relatively small, below .50. By contrast, for 30 and 50 clusters, power is relatively higher than .80. When examining the cluster size conditions, power is lower than .50 when the cluster size is 6, close to .80 when the cluster size 17, and higher than .80 when the cluster size is 32. A similar pattern is found across the three levels of ICC.

The power of γ_{20} is relatively higher, reaching values over .90 in the three distributions, the number of clusters, cluster size, and ICC. However, in the number of clusters, cluster size, and ICC conditions, power increases as the level of each condition increases. When examining the type I error rate for both γ_{10} and γ_{20}, their values are in the range of .036 to .063, suggesting that the Type I error rate is not upward or downward estimated.

4.3.2. Results by cells.

4.3.2.1. Relative bias γ_{10} and γ_{20}. Table 19 summarizes the relative (and absolute) bias of γ_{10}. The average cell bias across cells ranges from -0.044 to 0.023, which implies that γ_{10} deviates between ± 5% from the specified parameter value. This suggests that γ_{10} is not biased under any condition, but it may have variation due to other than random error. The interpretation of values in Table 19 is the same as in Table 11 in Experiment 1.

Table 19

γ_{10} Relative Bias for All Conditions in Experiment 2

	10 Clusters			30 Clusters			50 Clusters		
	Cluster size			Cluster size			Cluster size		
ICC	6	17	32	6	17	32	6	17	32
				Normal distribution					
.05	0.001	-0.038	0.015	0.002	0.023	-0.001	-0.018	-0.008	-0.008
.15	0.013	-0.010	0.008	-0.005	0.009	0.002	-0.004	0.002	0.002
.25	-0.026	-0.004	-0.009	-0.009	0.001	0.001	0.003	-0.013	-0.004
				t distribution 4 df					
.05	-0.087	-0.003	-0.044	0.001	0.022	0.005	0.011	-0.017	0.008
.15	0.010	-0.013	0.003	-0.026	-0.008	0.000	0.000	-0.001	-0.007
.25	-0.015	-0.023	-0.012	-0.014	-0.006	-0.002	0.006	-0.013	-0.003
				t distribution 11 df					
.05	-0.009	-0.003	-0.007	0.023	-0.019	0.005	-0.009	0.011	0.006
.15	0.001	0.009	0.004	-0.012	-0.015	-0.012	0.007	0.000	-0.003
.25	0.003	-0.003	-0.026	0.012	0.002	0.007	0.014	-0.002	0.001

To check whether statistically significant variation exists, an inferential analysis, via ANOVA, was performed. The results of the inferential analysis confirmed that none of the conditions produce a systematic variation in the values of γ_{10} ($\eta_p^2 < .09$). See Table 33 in the appendix.

Table 20 summarizes the relative bias of γ_{20}. The average bias across cells suggests that γ_{20} deviates between ± 5% from the specified parameter value. This implies that γ_{20} is not biased under any condition, but still may present statistically significant variation across conditions. To address this issue, an inferential analysis, via ANOVA, was conducted. The analysis showed statistically significant variation. This variation is present in a two-way interaction term between the cluster size and the ICC (η_p^2=.122) and in a three-way interaction between the number of clusters, the cluster size and ICC (η_p^2=.290). Their effect sizes are small and large, respectively (see Table 36 in the

appendix). The three-way interaction suggests that increasing the number of clusters is very powerful in terms of explaining patterns, i.e., the two-way interaction for relative bias involving cluster size and ICC varies sharply across different numbers of clusters. Figures 9 and 10 graphically represent these interactions.

Table 20

γ_{20} Relative Bias for All Conditions in Experiment 2

	10 Clusters			30 Clusters			50 Clusters		
	Cluster size			Cluster size			Cluster size		
ICC	6	17	32	6	17	32	6	17	32
				Normal distribution					
.05	-0.035	0.001	0.010	-0.013	0.000	0.001	0.008	-0.002	-0.002
.15	-0.004	0.002	-0.003	-0.001	0.001	0.000	0.005	0.002	0.003
.25	0.002	0.002	-0.002	0.003	-0.006	-0.001	0.000	0.000	0.000
				t distribution 4 df					
.05	-0.009	-0.016	0.004	-0.007	-0.004	-0.001	0.000	-0.003	0.000
.15	-0.004	-0.001	-0.002	-0.009	-0.001	-0.002	-0.001	-0.001	0.001
.25	0.008	0.007	-0.005	-0.006	-0.001	0.001	-0.004	-0.001	0.002
				t distribution 11 df					
.05	-0.024	-0.002	0.009	-0.019	-0.001	0.000	0.010	-0.001	-0.002
.15	0.001	-0.005	0.002	-0.001	-0.001	0.000	-0.002	-0.001	0.001
.25	0.008	0.001	0.000	-0.006	-0.003	0.000	-0.001	0.001	-0.001

Note: The interpretation of values in Table 20 is the same as in Table 11 in Experiment 1.

The two-way interaction suggests that the cluster size levels tended to produce different relative bias values of γ_{20} at the different levels of ICC, but these are still within the ± 5% tolerance range. Figure 15 shows that at 0.05 and 0.15 levels of ICC, the relative bias of γ_{20} across cluster size presents more variation, and γ_{20} is relatively inaccurate compared with the other ICC level. However, as ICC increases, the variation in the relative bias across cluster size levels is substantially reduced and approaches zero.

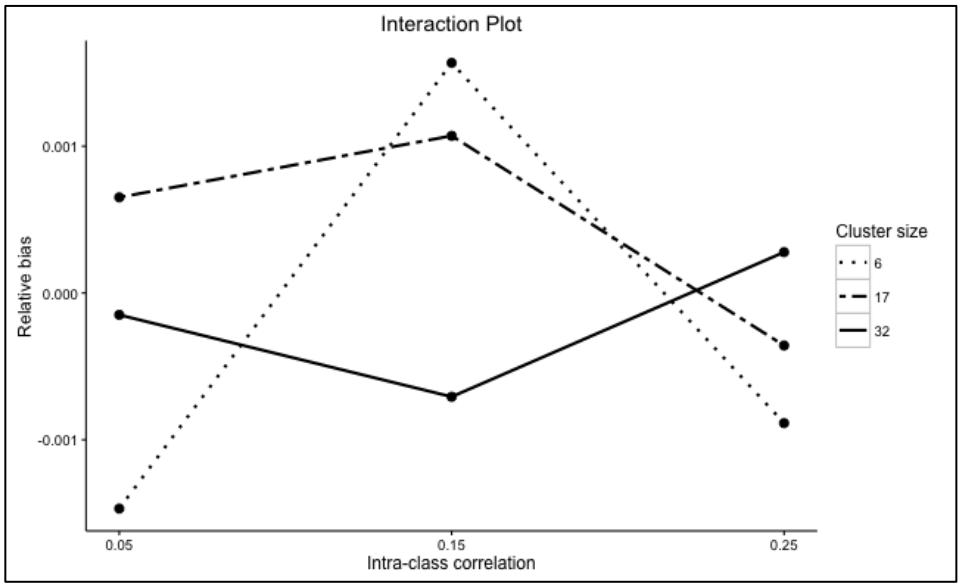

Figure 9. Corrected interaction effect between the cluster size and the ICC on γ_{20} relative bias.

The three-way interaction indicates that the interaction between cluster size and ICC produced different relative bias values across the levels of the number of clusters. Figure 10 shows this pattern. In general, at any level of the number of clusters, the cluster size and ICC interaction show higher variability in the relative bias of γ_{20} at the lowest ICC. However, when the ICC is .15, the variation in γ_{20} relative bias decreases substantially, and γ_{20} is estimated more accurately. When the level of ICC increases to .25, the relative bias of γ_{20} increases in variability, but it is smaller than when the ICC is .15.

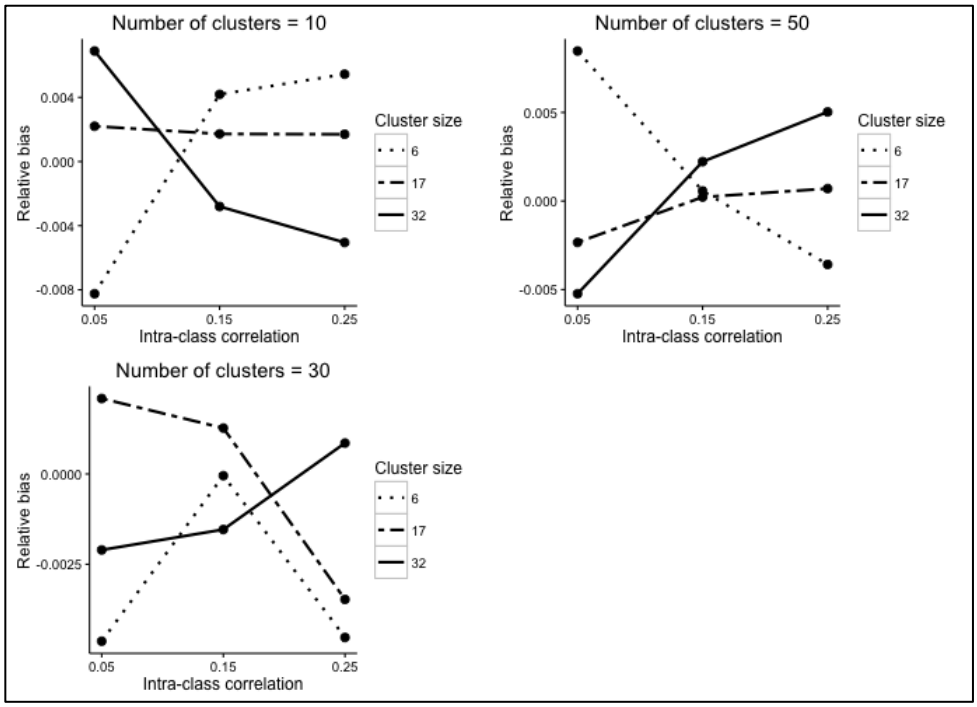

Figure 10. Corrected interactions effects on γ_{20} relative bias. Panel A shows the interaction between the cluster size and the ICC when the number of clusters is 10. Panel B shows the interaction between the cluster size and the ICC when the number of clusters is 30. Panel C shows the interaction between the cluster size and the ICC when the number of clusters is 50.

4.3.2.2. RMSE of γ_{10} and γ_{20}. The accuracy of γ_{10} was investigated by estimating the root mean square error (RMSE) as previously defined. Across cells, the minimum and the maximum values of RMSE are 0.124 and 0.957, respectively (see Table 21). These values are similar to those presented in Experiment 1. As before, the RMSE presents several patterns that can be visually identified. The RMSE decreases as the ICC, the cluster size, and the number of clusters increase. See Figure 11. This implies that the estimation of γ_{10} becomes more accurate as the ICC, the cluster size, and the number of clusters increase. In addition, when comparing the RMSE of the normal distribution with

the RMSE of the two t distributions, the values seem to increase as the level 2 error distribution departure from the normal distribution. This suggests that γ_{10} is more precisely estimated in the normal distribution.

Moreover, the RMSE values in the normal distribution fall below 0.30 (value arbitrarily selected) when the number of clusters are 30, the cluster size is 17, and the ICC is .15. A similar situation happens in the t distribution with eleven degrees of freedom. In contrast, in the t distribution with four degrees of freedom, this situation does not occur.

Table 21

RMSE of γ_{10} Across Conditions in Experiment 2

	10 Clusters			30 Clusters			50 Clusters		
	Cluster size			Cluster size			Cluster size		
ICC	6	17	32	6	17	32	6	17	32
				Normal distribution					
.05	0.847	0.570	0.477	0.485	0.315	0.269	0.372	0.261	0.207
.15	0.535	0.415	0.380	0.311	0.238	0.216	0.251	0.181	0.163
.25	0.447	0.381	0.348	0.262	0.211	0.207	0.199	0.162	0.154
				t distribution 4 df					
.05	1.201	0.825	0.643	0.686	0.458	0.384	0.538	0.366	0.290
.15	0.765	0.568	0.539	0.442	0.343	0.323	0.335	0.260	0.230
.25	0.619	0.533	0.484	0.359	0.307	0.275	0.274	0.236	0.219
				t distribution 11 df					
.05	0.943	0.628	0.526	0.564	0.359	0.288	0.413	0.288	0.228
.15	0.604	0.454	0.399	0.342	0.255	0.236	0.262	0.201	0.183
.25	0.494	0.418	0.380	0.294	0.238	0.220	0.224	0.178	0.171

To confirm the pattern, an inferential analysis was conducted. The WLS regression model performed for ln(RMSE) showed a negative effect for the number of clusters (β_2 = - 0.019, p-value < .01), cluster size (β_3 = - 0.023, p-value < .01) and ICC (β_4 = - 3.32, p-value < .01). The remaining variables did not have an impact on the ln(RMSE).

These results suggest that as the number of clusters, the cluster size and the ICC increase, the RMSE decreases in 0.019, 0.023 and 3.32 units, respectively (See Table 37). In other words, the fixed effect (γ_{10}) is more precisely estimated as the number of clusters, cluster size, and ICC increase.

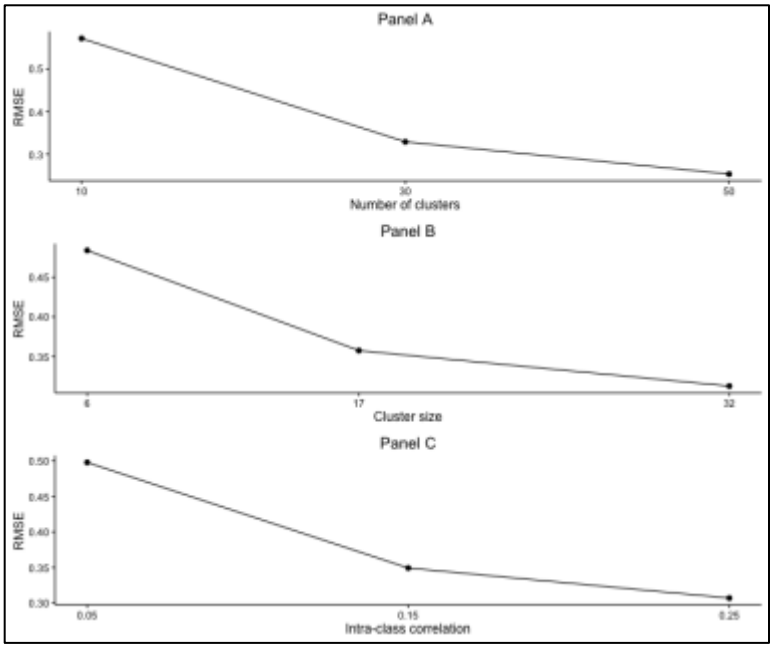

Figure 11. Main effects on γ_{10} RMSE. Panel A shows the main effect of the number of clusters on RMSE. Panel B shows the main effect of the cluster size on RMSE. Panel C shows the main effect of the ICC on RMSE.

The RMSE of γ_{20} was also investigated (see Table 22). Across cells, the minimum and the maximum values are .025 and 0.463, respectively. This range is narrow compared to the γ_{10} range, suggesting that γ_{20} is estimated more precisely than γ_{10}. In addition, the RMSE of γ_{20} presents the same patterns of the RMSE of γ_{10}. RMSE decreases as the ICC, the cluster size, and the number of clusters increase. See Figure 12. These patterns imply that the estimation of γ_{20} becomes more accurate as the ICC, the cluster size, and the

number of clusters increase. Moreover, when the RMSE of the normal distribution is compared with the RMSE of the two t distributions, the values are smaller in the normal distribution, suggesting that the γ_{20} is more precise in such a distribution.

Additionally, the RMSE values in the three distributions fall below .10 (value arbitrarily selected) when the number of clusters is 30, the cluster size is 17, and ICC is .15 or higher. A similar situation occurs in the t distribution with four and eleven degrees of freedom.

Table 22

RMSE of γ_{20} Across Conditions in Experiment 2

	10 Clusters			30 Clusters			50 Clusters		
	Cluster size			Cluster size			Cluster size		
ICC	6	17	32	6	17	32	6	17	32
				Normal distribution					
.05	0.425	0.236	0.170	0.229	0.138	0.098	0.178	0.105	0.080
.15	0.232	0.129	0.092	0.132	0.073	0.053	0.099	0.059	0.043
.25	0.167	0.091	0.070	0.095	0.055	0.040	0.073	0.042	0.031
				t distribution 4 df					
.05	0.578	0.340	0.236	0.322	0.194	0.139	0.260	0.150	0.111
.15	0.331	0.186	0.133	0.176	0.113	0.078	0.145	0.084	0.061
.25	0.232	0.135	0.100	0.134	0.077	0.056	0.103	0.061	0.045
				t distribution 11 df					
.05	0.439	0.260	0.188	0.259	0.150	0.108	0.198	0.119	0.087
.15	0.250	0.149	0.105	0.145	0.082	0.060	0.110	0.066	0.046
.25	0.179	0.110	0.077	0.104	0.062	0.043	0.080	0.049	0.034

The inferential analysis (WLS regression model) performed for γ_{20} ln(RMSE) suggests a negative effect for the number of clusters (β_2 = - 0.022, p-value < .01), cluster size (β_3 = - 0.036, p-value < .01) and ICC (β_4 = - 4.744, p-value < .01). The remaining variables did not have an impact on the ln(*RMSE*). These results suggest that as the

number of clusters, the cluster size and the ICC increase, the RMSE decreases in 0.022, 0.036 and 4.744 units, respectively (See Table 38). In other words, the fixed effect (γ_{10}) is more precisely estimated as the number of clusters, cluster size, and ICC increase.

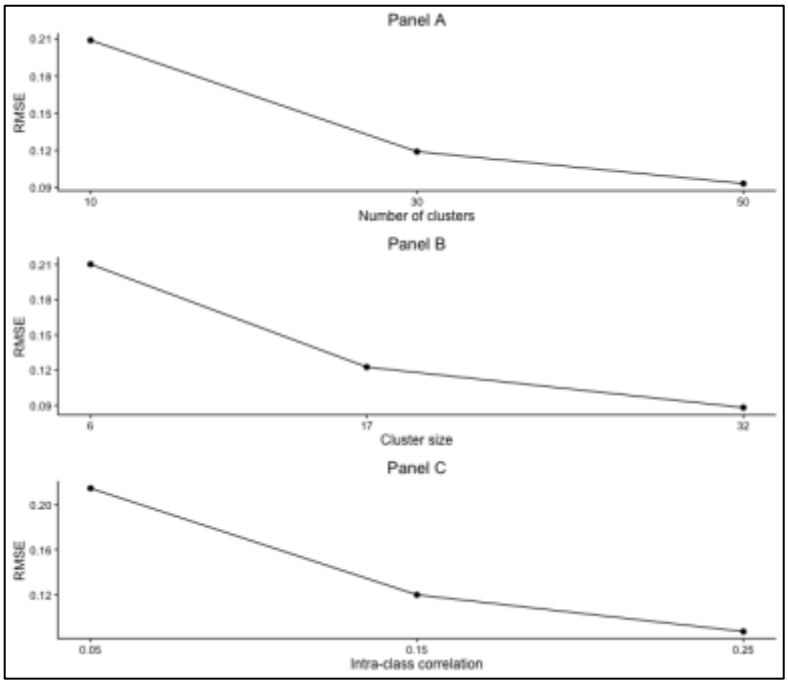

Figure 12. Main effects on γ_{20} RMSE. Panel A shows the main effect of the number of clusters on RMSE. Panel B shows the main effect of the cluster size on RMSE. Panel C shows the main effect of the ICC on RMSE.

Something important to note is that the distribution type did not have an impact on the accuracy of the fixed effects (γ_{10} and γ_{20}). This may suggest that estimating γ_{10} and γ_{20} among the error distributions in both experiments is almost equally precise. In other words, error distribution type does not impact the accuracy of the fixed effect estimation.

4.3.2.3. Power of γ_{10} and γ_{20}. Power by cells of γ_{10} is presented in Table 23. Power shows the same patterns as in Experiment 1. First, power increases as the number of clusters and the cluster size increase. Second, in the normal distribution a power of .80

(non-conservative value) is reached at 30 clusters with 17 observations within cluster. This happens when the ICC is .05. However, when the ICC is .15 or .25, a power of .80 is reached when at 30 clusters with 6 observations in each cluster. A similar pattern is found in the t distribution with eleven degrees of freedom. However, in this distribution a power value.80 is reached when the ICC is .05 and at 30 clusters with 32 observations in each cluster. However, when the ICC is .15 or .25, a power of .80 is reached at the same threshold as in the normal distribution. In the case of the t distribution with four degrees of freedom, a power of .80 is reached when the ICC value is .05 with 50 clusters and 32 observations within each cluster. However, when the ICC is .15 or .25, a power of .80 is reached at 30 clusters with 17 observations within each cluster. Finally, Table 23 shows that power increases as the ICC increases.

This pattern can be easily seen across the number of clusters and the cluster size. The reason for this positive relationship was explained in Chapter 3. Again, the patterns presented above suggest that the PNRCT model is underpowered, especially when the number of clusters is 10 and the cluster sizes are 6, 17 and 32. This also happens when the number of clusters are 30 and 50 and the cluster size is the lowest. Note that this situation is exacerbated when the level 2 error distributions are non-normal.

Table 23

Power of γ_{10} by Conditions in Experiment 2

	10 Clusters			30 Clusters			50 Clusters		
	Cluster size			Cluster size			Cluster size		
ICC	6	17	32	6	17	32	6	17	32
				Normal distribution					
.05	.187*	.377*	.517*	.521*	.865	.951	.720*	.976	.997
.15	.437*	.600*	.693*	.882	.985	.992	.985	1.000	1.000
.25	.521*	.680*	.747*	.960	.996	.999	.996	1.000	1.000
				t distribution 4 df					
.05	.095*	.230*	.285*	.293*	.612*	.754*	.463*	.786*	.928
.15	.248*	.406*	.494*	.598*	.838	.890	.839	.962	.980
.25	.323*	.456*	.508*	.766*	.886	.926	.953	.981	.982
				t distribution 11 df					
.05	.161*	.342*	.433*	.465*	.764*	.914	.649*	.942	.991
.15	.375*	.582*	.620*	.820	.958	.984	.962	.999	1.000
.25	.490*	.612*	.666*	0.938	.976	.990	.993	.999	1.000

Note: * denotes power lower than .80. The interpretation of power values in Tables 23 and 24 is the same as in Table 13.

To determine whether the variation across cells is random, an inferential analysis was conducted. The analysis indicates that three conditions have considerable variation. These conditions are the number of clusters, which has the largest effect size (η_p^2=.622), the ICC (η_p^2=.138), and the cluster size (η_p^2=.111), both with low effect size. It is important to notice that difference in variation was found across distributions, but its effect size was negligible ($\eta_p^2 < .09$) (see Tables 39 and 40 in the Appendix). The following figure illustrates the variation of the number of clusters, the cluster size and the ICC.

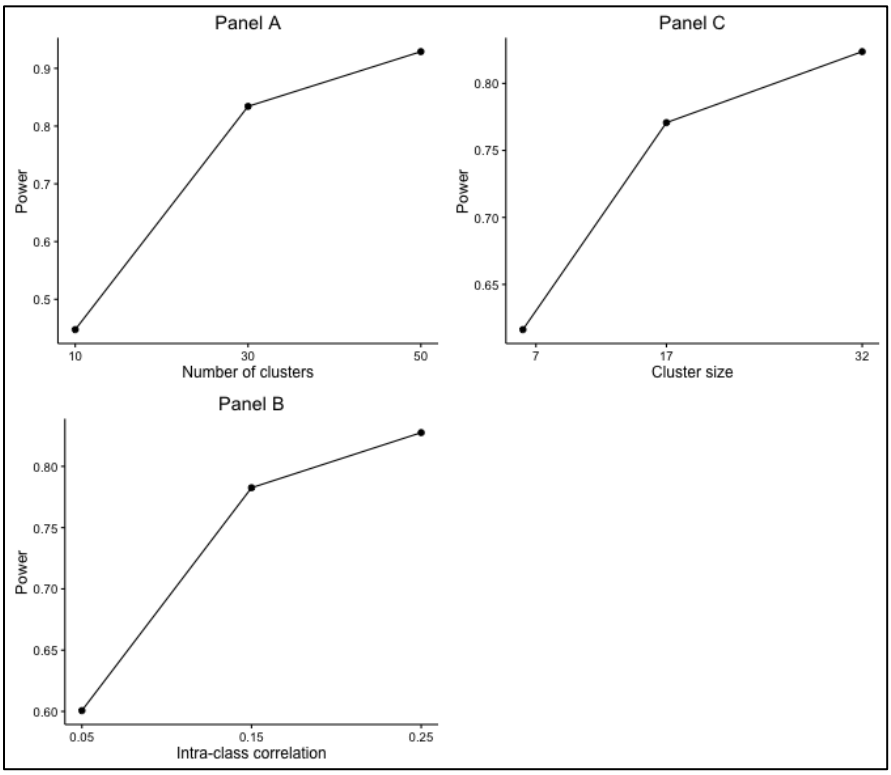

Figure 13. ANOVA conditions' main effects on γ_{10} power. Panel A shows the main effect of the number of clusters on power. Panel B shows the main effect of the ICC on power. Panel C shows the main effect of the cluster size on power.

The power of γ_{20} is presented in Table 24. This table shows that for the three distributions power is relatively low only when the number of clusters is 10, the ICC is .05 and the cluster size is 6. Power takes a value of 1 in other conditions. An important result to note in Table 24 is that the PNRCT model does not reduce the power of a covariate. Due to the large number of ones in the majority of the cells in Table 24, it is hard to visually identify a pattern. However, the inferential analysis shows statistically significant variation across the number of clusters (η_p^2=.103), cluster size (η_p^2=.099), and ICC (η_p^2=.110).

Additionally, statistically significant variation is found in the two- way interactions between the number of clusters and cluster size (η_p^2=.128), the two-way

interaction between the number of clusters and ICC (η_p^2=.145), and the two-way interaction between cluster size and ICC (η_p^2=.140). Moreover, the three-way interaction between the number of clusters, cluster size, and ICC (η_p^2=.167) shows statistically significant variation. All effect sizes range from low to medium (see Table 42 in the Appendix). Figures 14, 15, and 16 illustrate these interactions. It is important to notice that difference in variation was found across distributions, but its effect size was negligible (η_p^2 < .09) (see Table 41 and 42 in the Appendix).

Table 24

Power of γ_{20} by Conditions in Experiment 2

	10 Clusters			30 Clusters			50 Clusters		
	Cluster size			Cluster size			Cluster size		
ICC	6	17	32	6	17	32	6	17	32
				Normal distribution					
.05	.660*	.989	1.000	.985	1.000	1.000	1.000	1.000	1.000
.15	.988	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.25	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
				t distribution 4 df					
.05	.406*	.815	0.987	.857	1.000	1.000	.972	1.000	1.000
.15	.862	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.25	.994	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
				t distribution 11 df					
.05	.573*	.966	1.000	.965	1.000	1.000	.999	1.000	1.000
.15	.975	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
.25	.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Note: * denotes power lower than .80. The interpretation of power values in Tables 23 and 24 is the same as in Table 13.

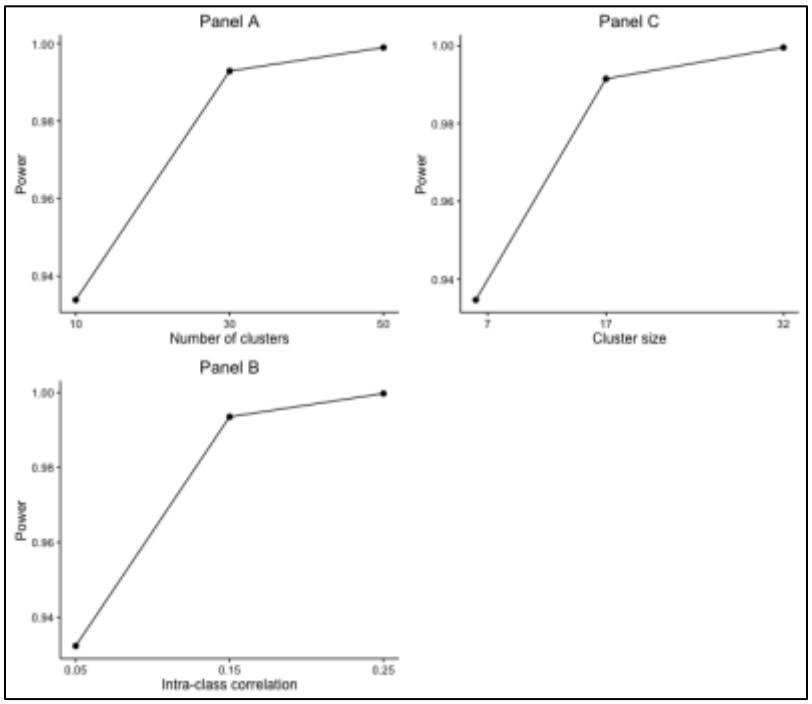

Figure 14. ANOVA conditions' main effects of γ_{20} power. Panel A shows the main effect of the number of clusters on power. Panel B shows the main effect of the ICC on power. Panel C shows the main effect of the cluster size on power.

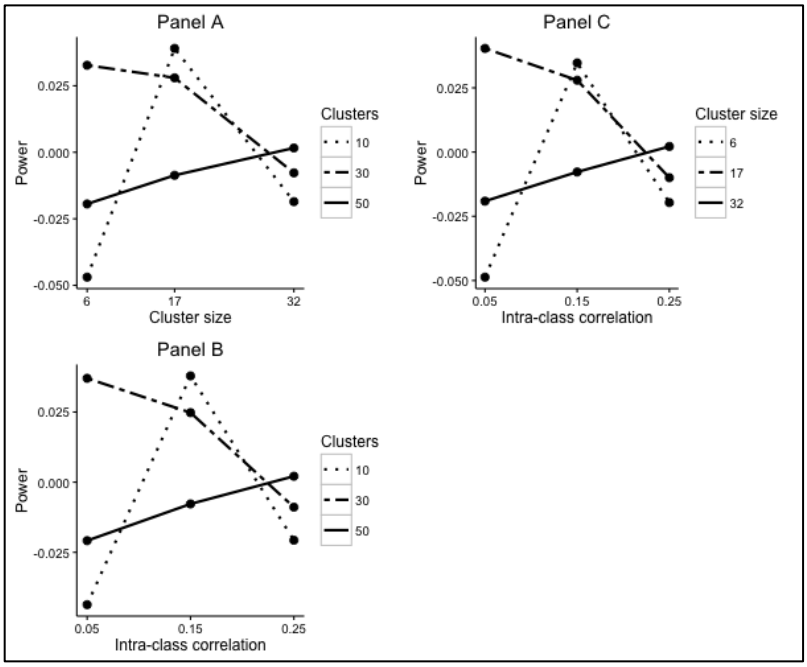

Figure 15. Corrected interaction effects of γ_{20} power. Panel A shows the interaction between the number of clusters and the cluster size on power. Panel B shows the interaction between the number of clusters and the ICC on power. Panel C shows the interaction between the cluster size and the ICC on power.

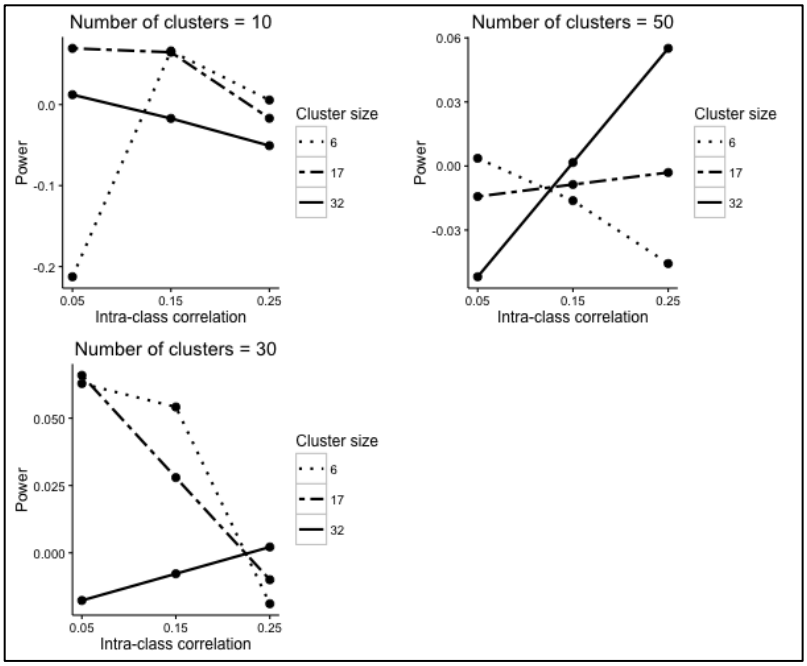

Figure 16. Corrected interactions effects for γ_{20} relative bias. Panel A shows the interaction between the cluster size and the ICC when the number of clusters is 10. Panel B shows the interaction between the cluster size and the ICC when the number of clusters is 30. Panel C shows the interaction between the cluster size and the ICC when the number of clusters is 50.

For the two-way interaction between the number of clusters and cluster size, power differs substantially across clusters at the lowest level of cluster size. However, as the cluster size increases, power seems to be more similar across clusters. Power increases as both the number of clusters and cluster size increase (see Figure 15, panel A). The two-way interaction between the number of clusters and ICC shows that as ICC increases, power is more similar across the number of clusters (see Figure 15, panel B). This pattern is the same in the two-way interaction between cluster size and ICC (see Figure 15, panel C). A similar situation happens in the three-way interaction between number of clusters, cluster size, and ICC (Figure 16). Power shows different values at the number of clusters and cluster size levels, across the levels of ICC. This pattern shows that power is more similar across cluster size as ICC increases. This is similar for 10 and 30 clusters. However, for 50 clusters the pattern is erratic.

4.3.2.4. Type I error rate of γ_{10} and γ_{20}. The values of the Type I error rate of γ_{10} range from .036 to .070. Most of the values are neither upward or downward estimated. However, two cells are upward estimated. The first of these two cells is in the condition of normal distribution with 10 clusters, a cluster size of 32, and an ICC of .15. The second cell is located in the t distribution with four degrees of freedom, 10 clusters, a cluster size of 17, and an ICC of .05.

The interpretation of the Type I error rate values in Table 25 is the same as suggested for Table 17 in Experiment 1.

Table 25

Type I Error Rate of γ_{10} by Conditions, in Experiment 2

	10 Clusters			30 Clusters			50 Clusters		
	WCO			WCO			WCO		
ICC	6	17	32	6	17	32	6	17	32
				Normal distribution					
.05	0.047	0.051	0.043	0.046	0.036	0.053	0.043	0.052	0.048
.15	0.046	0.058	0.070*	0.043	0.045	0.059	0.058	0.048	0.052
.25	0.056	0.056	0.056	0.062	0.041	0.057	0.051	0.046	0.049
				t distribution 4 df					
.05	0.039	0.069*	0.043	0.042	0.044	0.059	0.047	0.054	0.041
.15	0.045	0.056	0.044	0.055	0.061	0.057	0.045	0.049	0.050
.25	0.047	0.065	0.039	0.044	0.057	0.046	0.044	0.056	0.041
				t distribution 11 df					
.05	0.038	0.050	0.058	0.053	0.041	0.041	0.040	0.054	0.049
.15	0.050	0.047	0.060	0.051	0.039	0.055	0.050	0.051	0.055
.25	0.053	0.057	0.049	0.050	0.056	0.045	0.047	0.046	0.047

Note: * indicates Type I error rate downward or upward estimated.

Due to the fact that the Type I error rate values range from .036 to .070, some variation due to other than random error may exist. Thus, an inferential analysis was conducted. This analysis suggests that γ_{10} presents statistically significant variation in the two-way interaction between distributions and cluster size (η_p^2=.230) and in the two-way interaction between number of clusters and cluster size (η_p^2=.141).

The first of these two interactions has a large effect size, and this implies that average Type I error rate values are different across distributions for each level of the cluster size. In this respect, as seen in Figure 17, panel A shows that the Type I error is different for each distribution type at each cluster size level. Figure 17, panel B shows the second interaction. This has a medium effect size, and this interaction suggests that the Type I error rates across the levels of the number of clusters are different. Note that this variation, for both two-way interactions, is still within the acceptable interval of the Type I error rate (from .036 to .063).

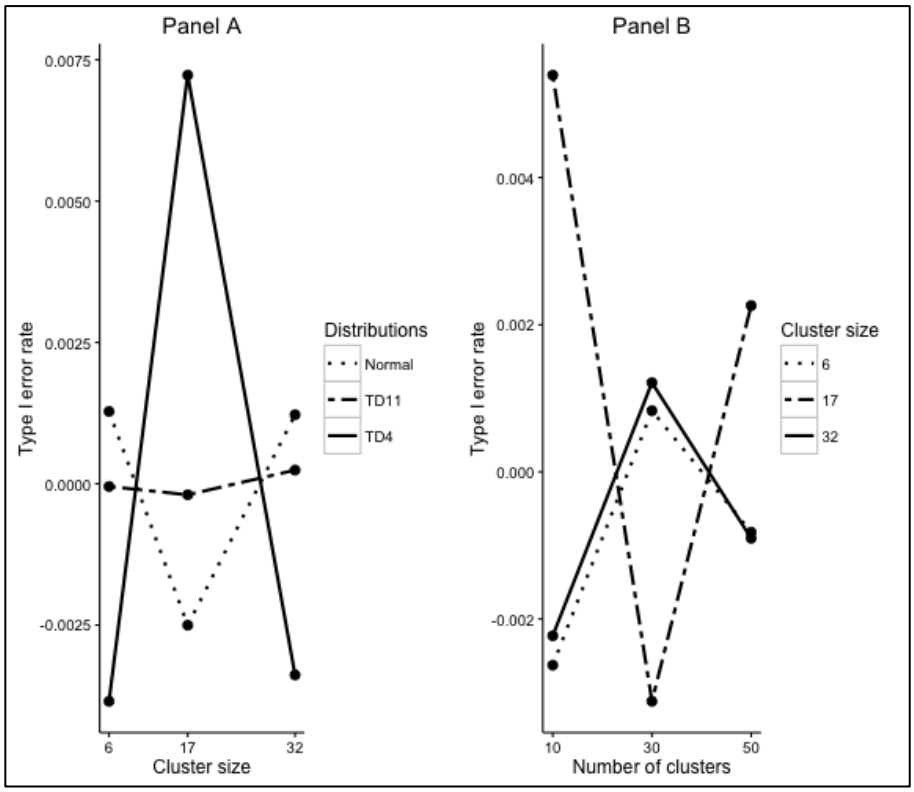

Figure 17. Corrected interaction effects on γ_{10} Type I error rate. Panel A shows the interaction between the cluster size and distributions on Type I error rate. Panel B shows the interaction between the cluster size and the number of clusters on Type I error rate.

The type I error rate values of γ_{20} range from .037 to .064, indicating that all values are not upward or downward estimated (see Table 26). The interpretation of the Type I error rate values in Table 26 is the same as suggested for Table 17 in Experiment 1.

The variation across conditions was examined by performing the respective inferential analysis. This analysis suggests that the Type I error rate of γ_{20} presents statistically significant variation between distributions (η_p^2=.120). The effect size of this variation is considered low according to the settings in Chapter 3 ($.09 \leq \eta_p^2 <.14$ is a low effect size; see Table 46 in the Appendix).

This statistically significant variation may be due to the Type I error rate values of the normal distribution. This Type I error rate values have a large deviation from the nominal Type I error compared to the two t distributions, which are close to the nominal Type I error rate values. Figure 18 shows the variation of the main effects between distributions.

Table 26

Type I Error Rate of γ_{20} by Conditions in Experiment 2

	10 Clusters			30 Clusters			50 Clusters		
	Cluster size			Cluster size			Cluster size		
ICC	6	17	32	6	17	32	6	17	32
				Normal distribution					
.05	0.050	0.048	0.046	0.045	0.052	0.037	0.043	0.040	0.050
.15	0.048	0.055	0.040	0.053	0.037	0.037	0.054	0.055	0.045
.25	0.043	0.042	0.053	0.050	0.055	0.051	0.045	0.045	0.043
				t distribution 4 df					
.05	0.057	0.046	0.041	0.047	0.051	0.049	0.054	0.046	0.053
.15	0.063	0.050	0.045	0.040	0.062	0.055	0.049	0.051	0.053
.25	0.053	0.056	0.051	0.050	0.048	0.038	0.049	0.048	0.049
				t distribution 11 df					
.05	0.045	0.052	0.049	0.054	0.050	0.041	0.046	0.051	0.052
.15	0.057	0.061	0.052	0.055	0.046	0.049	0.050	0.061	0.051
.25	0.059	0.060	0.057	0.046	0.046	0.044	0.055	0.059	0.048

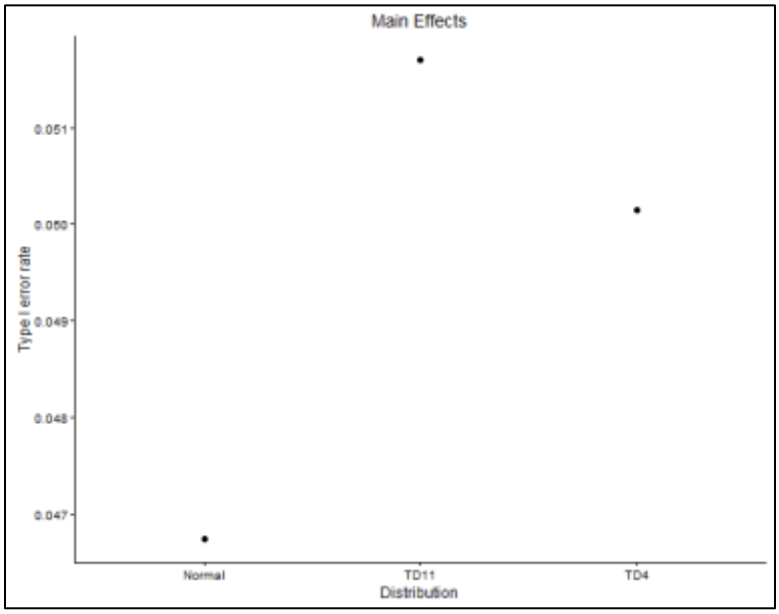

Figure 18. ANOVA conditions main effects of the distributions on γ_{20} Type I error rate.

In this figure, it can be seen that the most non-normal distribution (t distribution with 4 degrees of freedom) seems to produce, on average, a Type I error rate close to the nominal rate (.05). However, the Type I error rates produced by the t distribution with 11 degrees of freedom and the normal distribution are within their sampling error.

Chapter 5

Discussion and Limitations

Many statistical models assume normality of the error distribution. Violation of this model assumption may negatively affect the reliability and validity of model outcomes. The PNRCT model also assumes normality of the error distribution. Because this assumption may be unrealistic when working with real world data, it is important to understand the consequences and implications for the model outputs caused by violations to the normality assumption of the error distribution.

The current Monte Carlo study explored the implications in the PNRCT model when the normality assumption of the error distribution at level 2 does not hold. More specifically, the study tried to answer the following research questions:

> Are the fixed parameter estimates (γ_{10} and γ_{20}), Type I error rates, and power affected by (a) level 2 heavy-tailed error distribution, (b) cluster size, (c) number of clusters, and (d) levels of ICC? If so, to what extent?

To answer these questions, four conditions were explored, including three error distributions at level 2 of a PNRCT model with one covariate at level 1 (normal distribution, t distribution with four degrees of freedom, and t distribution with eleven degrees of freedom), three levels of cluster size (6, 17, 32), three levels of ICC (.05, .15, .25), and three levels of the number of clusters (10, 30, 50).

This chapter discusses the results and the associated conclusions particularly as they impact statistical practice, reviews limitations, and offers recommendations for future research.

5.1. Discussion

One aspect the research questions addressed was whether the fixed parameter estimates, γ_{10} and $\gamma_{20,}$ were affected by the conditions of the experiment, including non-normal error distributions at level 2. The results of the present study showed that the fixed effects on average were for the most part unbiased (the range was within the region of tolerance) and none of the error distributions showed significant variation in the relative bias of the fixed effects for either of the two experiments, the pure PNRCT model and the PNRCT model adjusted by one covariate.

These results are not uncommon, as researchers have reported that fixed effects are not biased when the normality assumption is violated in hierarchical non-PNRCT models under similar conditions to those presented in this study (e.g., Ketelsen, 2014; Mass & Hox, 2005a). To the best of my knowledge, researchers have not offered any explanation for why the fixed effects are not biased (e.g., Kim, 1991; Raudenbush & Bryk, 2002). They usually claim that under non-normality the fixed effects are expected to remain unaffected (e.g., Mass & Hox, 2005a). However, fixed effects are unbiased or slightly biased because they represent conditional means and they are subject to the central limit theorem.

None of the other conditions of the study (number of clusters, cluster size, or ICC) showed any pattern of bias on the fixed parameter estimates. Other researchers' findings showed similar conclusions (e.g., Shieh, 1999; Shieh, Fouladi, & Pullum, 2001). In my results, the fixed effects were robust even when the number of clusters and cluster size in the treatment group were relatively small (ten clusters and six observations). The fact that cluster size and the number of clusters did not impact the bias of the fixed effects is

particularly important. This may support the use of the PNRCT model with small numbers of clusters and clusters sizes. For instance, researchers who have a small budget and whose research design fits the PNRCT model, but who do not care at all about power, may use 10 clusters in the treatment group with 6 subjects within groups, and 60 subjects in the control group with confidence that the average treatment effect will not be biased. This suggests that a research study with PNRCT design may be conducted with 120 subjects, which results in a low-cost study. However, researchers have to be aware that with this sample size power will be very low (around .20).

Another possibility for researchers with small budgets is to increase somewhat the sample size. Researchers may use 30 clusters with 6 observations, which makes the treatment group sample size 180 individuals, so the sample size for the control groups will also be 180 individuals. In total, the sample size will be 360 individuals. The fixed effect will still be unbiased by using this sampling plan, but power may increase up to 0.96, depending on the ICC.

The results also support the use of the PNRCT model, as a research design, for other researchers interested in the average treatment effect, and with the possibility of using a large number of clusters (up to 50 clusters) and a large number of subjects within each cluster (32 subjects). With this condition, the treatment group sample size is 1600 individuals, with the same number of individuals in the control group. The total sample size will be 3200 subjects.

In contrast to Shieh, Fouladi and Pullum (2001) and Ketelsen (2015) the results of this study show that the fixed effects were not biased for any of the ICC levels. The reasons the results of this study differ from other studies for ICC may be twofold. First,

other studies have found an impact of ICC on fixed effects, but at higher levels than those presented in this study. For instance, Shieh, Fouladi and Pullum (2001) found that an ICC level of 0.5 biased the fixed effects. Second, in other studies the ICC includes both treatment and control groups, but in the PNRCT models the ICC only applies to the treatment group.

In general, the fact that the fixed effects of the PNRCT in Model A and Model B present almost no bias has positive implications for researchers who used or pretend to use the PNRCT model in applied settings. More specifically, those researchers with a research design with similar conditions to those used in the present research do not have to worry about whether the estimation produces bias in the treatment effect because the likelihood of obtaining unbiased or negligible bias in the treatment effect will be high.

The results of my study also showed that in both Experiment 1 and Experiment 2 the distribution type did not have an impact on the accuracy (RMSE) of the fixed effects (γ_{10} and γ_{20}). This suggests, for the conditions of the study (symmetric heavy tail level 2 error distributions), that the error distribution type does not impact the accuracy of the fixed effect estimation. I did not expect the distribution type to have any impact on the variability because accuracy is related to other factors such as the sample size, which includes the number of clusters and the cluster size. In this regard, in both Experiment 1 and 2, the fixed effects were more precisely estimated as the number of clusters and cluster size increased. In other words, the variability was reduced as the total sample size increased. This is consistent with statistics theory, which states that the sample size is always a factor that increases the accuracy of parameter estimates. The total sample size

in this research is a factor of the number of clusters and the cluster size (e.g., 10 clusters and 6 cluster sizes made a total sample size of 60). When the number of clusters was fixed (e.g., 10) but had different cluster sizes (6, 17, and 32), the total sample size of the treatment group increased (60, 170, and 320). Similarly, when the number of clusters increased (10, 30, and 50) but the cluster size was fixed (e.g., 6) the total sample size of the treatment group also increased (60, 180, and 300). Therefore, the reason why the accuracy of the fixed effect estimation increased as both the number of clusters and the cluster sizes increased is because the total sample size increased.

Both Experiment 1 and 2 showed that the fixed effects were also more accurately estimated as the ICC increased. This situation is not uncommon in simulation studies that involve hierarchical data structure. Researchers have reported that the fixed effects are more precisely estimated as the ICC increases (Baldwin et al., 2011; Max & Hox, 2001; Snijders & Bosker, 1994, 1999; Shieh & Fouladi, 2003). I argue that as the ICC increases, the variation between clusters is increased, so when more variation is present, the accuracy of parameter estimates increases.

This result implies that researchers who plan to use a PNRCT model and who also are interested in obtaining parameter estimates with less dispersion (more accuracy) should use a large sample size in their design. They can use 50 clusters and 32 subjects within clusters in the treatment group (n_1 = 1600) and 1600 subjects in the control group (n_2). This makes a total sample size of 3200 subjects. There is nothing to do regarding the ICC because the ICC in applied settings cannot be controlled; it is calculated with the variance components.

The main research question of this study also addressed whether the power of the fixed effects was impacted by the conditions of the study. The ANOVA analysis suggested that the distribution type was statistically significant regarding the power of γ_{10}, in both Experiment 1 and Experiment 2, and regarding the power of γ_{20} in Experiment 2. However, the distribution type had a negligible effect size. In fact, the results showed that the PNRCT model had an under-power of γ_{10} under the normality assumption, and departures from the normal error distribution at level 2 might exacerbate the under-power of γ_{10}. This is because power was reduced when the level 2 error distribution was t distributed with eleven degrees of freedom. The situation was worse when the level 2 error distribution was t distributed with four degrees of freedom. This mainly happened across cluster sizes when the numbers of clusters were 10 and 30. However, when the number of clusters was 50 this situation occurred at cluster sizes of 6 and 17. In the case of γ_{20}, the problem was critical when the number of clusters was 10 and the cluster size was 6.

In addition, the results showed that power increased as the number of clusters and the cluster size increased. The effect size of these conditions showed statistically significant variations in both experiments. These results were consistent with other studies which indicated that power was positively affected by the number of clusters and the cluster size (e.g., Ketelsen, 2014; Kim, 1990; Kreft, 1996; Mass & Hox 2004a; Shih, 2008; Snijders, 2005). Snijders (2006) asserted that power is related to sample size at different levels in HLM. If this is true, then it follows that not only the number of clusters and the cluster size increase power, but also the interaction of these conditions.

These results imply that, if researchers are concerned about power, they must avoid working with a PNRCT design with a small number of clusters and any cluster size presented in this study. This is because under such conditions power will be far below .8 and it will decrease even more if the normal level 2 error distribution is not achieved. The same situation may be true for 30 and 50 clusters but only when the ICC is .05. Researchers who plan to use the PNRCT model and want to detect very small average treatment effects can use a large total sample size of 3200 subjects disaggregated in 50 clusters, with 32 subjects in the treatment group and 1600 subjects in the control groups. With this total sample size power may reach values that approach 1 even when the error distribution deviates from normality.

Some researchers may be liberal when thinking about power, so they could reduce the total sample. In the treatment condition, they could use a sample size of 510 subjects composed of 30 clusters, with 17 subjects within each cluster. The respective control group then would have 510 subjects. These values may produce a power higher than 0.8 even when the level 2 error distributions deviate from normality. Other researchers may be interested in small sample size but in relatively high power. They have two options. First, they could use a sample size of 320 subjects in the treatment group, composed of 10 clusters with 32 subjects within subjects and 320 subjects in the control group. Second, researchers could use a sample size of 180 subjects in the treatment group composed of 30 clusters with 6 subjects within each cluster and 180 subjects in the control group. In the first case, power may be up to 0.75, while in the second case power could assume values larger than .80. However, in both cases researchers have to rely on the ICC. In the

first case, researchers will need an ICC of 0.25, but in the second case, researchers will need an ICC value of 0.15 or .25.

The ICC also showed an impact on the power of the fixed effects in both experiments. The pattern suggested that the power of the fixed effect increased as the ICC increased. These results contradict previous studies which suggested that power decreases as ICC increases (e.g., Ketelsen, 2014; Shih, 2012). However, the positive effect of ICC on power in the present research is due to the way ICC was controlled in the experiments. This was explained in Chapter 3.

Overall, the results regarding power imply that data analysts and researchers need not worry too much when analyzing data from a research design with a large number of clusters, but they have to be concerned if they work with sizes 10 or 30 clusters with 6, 17 or 32 subjects.

A final aspect of the research question of this study was whether the Type I error rate of the fixed effects was affected by the conditions of the study. In both experiments the observed Type I error rates of γ_{10} and γ_{20} were not affected by the violation of the normality assumption at level 2. These results are supported by some studies that report acceptable (neither inflated nor deflated) Type I error rates (e.g., Ketelsen, 2014) when the assumption of normality is violated. In Experiment 2, although the Type I error rate was neither deflated or inflated, the inferential analysis indicated variations across distributions, but this variation was in the acceptable range. This may have happened because Model B in Experiment 2 was adjusted by a covariate. The insertion of this covariate in the model may have caused more accuracy in the estimates, so that the

analysis showed variation of the Type I error rate across distributions. Second, the large sample size probably also contributed to the inferential analyses detecting variations within Type I error rates.

These results imply that academics with a research design that fits the PNRCT model, with conditions similar to the present study, need not worry about the nominal Type I error. The chance of finding inflated or deflated Type I error rates even when the normality assumption is violated in the error distribution at level 2 will be very low. In addition, researchers should be confident that under conditions where the number of clusters, cluster size, or even ICC are similar to the present study, the Type I error will deviate only slightly from the nominal Type I error rate of .05.

The following table summarizes the results when the level 2 error distribution is non-normal; more specifically, when the level 2 error is t distributed.

Table 27

Summary of the Consequences of Violation Assumption of Normality in the Level 2 Error Distribution

	Model A				Model B				
Distribution	Bias of the Fixed Effects (Treatment Effect)	RMSE	Type I Error Rate	Power	Bias of the Fixed Effects (Treatment Effect)	Bias of the Fixed Effects (Covariate Effect)	RMSE	Type I Error Rate	Power
t_{11df}	Negligible effect	Negligible effect	Modest inflated and deflated	t test is underpowered	Negligible effect	Negligible effect	Negligible effect	Modest inflated and deflated	t test is underpowered
t_{4df}	Negligible effect	Negligible effect	Modest inflated and deflated	t test is underpowered	Negligible effect	Negligible effect	Negligible effect	Modest inflated and deflated	t test is underpowered

Note: RMSE in model B presented the same effect (negligible effect) on both treatment effect and covariate effect.

5.2. Limitations of this Study and Recommendations for Future Research

This study has some important limitations. First, the simulation conditions included only the t distribution with four and eleven degrees of freedom. Working with real world data, the level 2 PNRCT models may present other heavy-tailed distributions such as Chi-squared distributions with 1 and 4 degrees of freedom. Therefore, future research on PNRCT models, specifically on the violation of the normal assumption of error terms, can be pursued by evaluating non-symmetric heavy-tailed distributions. Since the PNRCT model outcomes and its estimation seem to be robust under non-normal heavy-tailed symmetric error distributions (t distributions with four and t distributions with eleven degrees of freedom), it's possible that non-symmetric heavy-tailed distributions (e.g., Chi-squared distributions with one and four degrees of freedom, or uniform distributions) would similarly not bias the fixed effects or impact the Type I error rates. However, this needs to be investigated in future studies, along with the impact of non-normality on estimated variance components in PNRCT.

Another important limitation is that this study used 10 clusters as the minimum number of clusters for the PNRCT models. However, in applied settings some researchers may avoid using the PNRCT model because the number of clusters is less than 10. For instance, Roberts et al. (2011) implemented a design that clearly fits the PNRCT model, but they used an ANOVA model. In this design, the treatment condition had 29 participants ($n_1 = 29$) and the control condition 27 participants ($n_2=27$). Participants in the treatment condition were assigned to clusters of five to six individuals, which implies four clusters of six individuals and one cluster of five individuals, in total 5 clusters.

In addition to this issue, in simulation studies the PNRCT model has been fitted with less than 10 clusters. For instance, Baldwin et al. (2011) used 2, 6 and 8 clusters; Tesller (2014) used 8 clusters; and Sanders (2011) used 2, 4, 5, and 8 clusters. These situations raise two issues. First, there is no evidence to prevent or encourage researchers to use the PNRCT model with less than 10 clusters other than the power issue. Second, it may be possible that by violating the normality assumption the PNRCT outcomes under this condition (number of clusters less than 10) may present different results, specially those related to biased and Type I error rate. Thus, future research could explore whether including numbers of clusters from 2 to 10, and varying the levels of cluster sizes (e.g., six, ten, twenty observations within each cluster), would impact the model outcomes. This may encourage or prevent, depending on the results, other researchers using the PNRCT model with less than 10 clusters.

Another limitation is that in this research, the level 1 error distribution remained normal for each level 2 error distribution. This assumption may be unrealistic in applied settings in which the level 2 error distribution is t-distributed and the level 1 error distribution remains normal. Future research studies may include conditions in which researchers violate the normality assumption at both levels.

A final limitation is that the PNRCT models used in this research were fitted under the assumption of equal variances for the clusters in the treatment condition. Therefore, it is possible that under variance heterogeneity at level 1 and non-normal residual distributions at level 2, the model outcomes may be negatively impacted. This may be a topic for future investigation.

References

Bainbridge, T. R. (1985). The committee on standards-precision and bias. *ASTM Standardization News*, *13*(1), 44-46. Retrieved from http://www.icast.org.in/ucat/csirill.php

Baldwin, S. A., Bauer, D. J., Stice, E., & Rohde, P. (2011). Evaluating models for partially clustered designs. *Psychological Methods*, *16*(2), 149. doi: 10.1037/a0023464

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1 - 48. doi: 10.18637/jss.v067.i01

Bauer, D. J., Sterba, S. K., & Hallfors, D. D. (2008). Evaluating group-based interventions when control participants are ungrouped. *Multivariate Behavioral Research*, *43*(2), 210-236. doi:10.1080/00273170802034810

Bloom, H. S., Bos, J. M., & Lee, S. W. (1999). Using cluster random assignment to measure program impacts: Statistical implications for the evaluation of education programs. *Evaluation Review*, *23*(4), 445-469. doi: 10.1177/0193841X9902300405

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis*, *29*(1), 30-59. doi: 10.3102/0162373707299550

Boruch, R., de Moya, D., & Snyder, B. (2002). The importance of randomized field trials in education and related areas. In F. Mosteler & R. Boruch (Eds.), *Evidence matters: Randomized trials in education research* (pp. 50-79). Washington, D.C.: Brookings Institution Press.

Bray, M. A., & Kehle, T. J. (Eds.). (2013). *The Oxford handbook of school psychology*. Oxford University Press.

Browne, W. J. (1998). *Applying mcmc methods to multi-level models* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (301570982)

Browne, W., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics*, *15*(3), 391-420. doi: 10.1007/s001800000041

Busing, F. M. (1993). *Distribution characteristics of variance estimates in two-level models: A Monte Carlo study*. Department of Psychology: University of Leiden. Netherlands.

Burstein, L., Kim, K. S., & Delandshere, G. (1989). Multilevel investigations of systematically varying slopes: Issues, alternatives, and consequences. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 233-276). San Diego, CA: Academic Press.

Burtless, G. (2002). Randomized field trials for policy evaluation: Why not in education? In F. Mosteler & R. Boruch (Eds.), *Evidence matters: Randomized trials in education research* (pp. 179-197). Washington, D.C.: Brookings Institution Press.

Caliendo, M., & SpringerLink. (2006). *Microeconometric evaluation of labour market policies.* Personal Collection of (Lecture notes in economics and mathematical systems; 568), Frankfurt University, Berlin.

Candel, M. J., & Van Breukelen, G. J. (2009). Varying cluster sizes in trials with clusters in one treatment arm: Sample size adjustments when testing treatment effects with linear mixed models. *Statistics in Medicine*, *28*(18), 2307-2324. doi**:**10.1002/sim.3620

Cerulli, G. (2015). *Econometric evaluation of socio-economic programs: Theory and applications* Advanced studies in theoretical and applied econometrics series 49. Heidelberg: Springer.

Coalition for Evidence-Based Policy. (2007). *When is it possible to conduct a randomized controlled trial in education at reduced cost, using existing data sources?* Retrieved from *http://coalition4evidence.org/468-2/publications/*

Cohen, J. (1969). *Statistical power analysis for the behavioural sciences*. New York, NY: Academic Press.

Coleman, J. L. (2006). *A simulation study of the piecewise hierarchical model approach to meta-analysis of single -subject data* (Doctoral dissertation). Retrieved from ProQuest. (305306229)

Compas, B. E., Forehand, R., Thigpen, J. C., Keller, G., Hardcastle, E. J., Cole, D. A., . . . Roberts, L. (2011). Family group cognitive-behavioral preventive intervention for families of depressed parents: 18- and 24-month outcomes. *Journal of Consulting and Clinical Psychology, 79*(4), 488-499. doi: 10.1037/a0024254

Coverdale, J. H., Balon, R., Beresin, E. V., Louie, A. K., Tait, G. R., & Roberts, L. W. (2013). An argument for conducting methodologically strong, randomized, controlled trials in educational research. *Academic Psychiatry*, *37*(3), 145-9. doi**:**10.1176/appi.ap.13030029

Darandari, E. Z. M. (2004). *Robustness of hierarchical linear model parameter estimates under violations of second -level residual homoskedasticity and independence assumptions* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (305182787)

Daniel, W. (2005). *Biostatistics : A foundation for analysis in the health sciences* (8th ed.). Wiley series in probability and statistics. Hoboken, NJ: Wiley.

Delpish, A. N. (2006). *A comparison of estimators in hierarchical linear modeling: Restricted maximum likelihood versus bootstrap via minimum norm quadratic unbiased estimators* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (305334243)

DiPrete, T. A., & Forristal, J. D. (1994). Multilevel models: Methods and substance. *Annual reviews of sociology, 20*, 331-357. doi: 10.1146/annurev.so.20.080194.001555

Fox, J. (1991). *Regression diagnostics* (Quantitative applications in the social sciences ; 79). Newbury Park, CA: Sage Publications.

Fox, J. (2008). *Applied regression analysis and generalized linear models*. Thousand Oaks, CA: Sage Publications.

Gail, M. H., Wieand, S., & Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, *71*(3), 431-444. doi: 10.2307/2336553

Gamst, G., Meyers, L. S., & Guarino, A. J. (2008). *Analysis of variance designs: A conceptual and computational approach with SPSS and SAS*. Cambridge; New York: Cambridge University Press.

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. (2011). *Impact evaluation in practice*. Washington, DC: World Bank Publications.

Glass, G., Peckham, P., & Sanders, J. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42*(3), 237-288. doi: 10.3102/00346543042003237

Goldstein, H. (1995). *Multilevel statistical models*. (2nd ed.). New York, NY: John Wiley.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association, 72*(358), 320-338. doi: 10.1080/01621459.1977.10480998

Harwell, M. (1998). Misinterpreting interaction effects in analysis of variance. *Measurement and Evaluation in Counseling and Development,31*(2), 125-36.

Harwell, M. & Kohli, N. (2015). Research design and data analysis in Monte Carlo studies. Manuscript submitted for publication.

Harwell, M. R., & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research, 71*,105-131. doi: 10.3102/00346543071001105

Hauck, W., & Anderson, S. (1984). A survey regarding the reporting of simulation studies. *The American Statistician, 38*(3), 214-216. doi: 10.1080/00031305.1984.10483206

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, 32, 341–370. doi:10.3102/1076998606298043

Hedges, L. V. (2011). Effect sizes in three-level cluster-randomized experiments. *Journal of Educational and Behavioral Statistics*, 36, 346–380. doi:10.3102/1076998610376617

Hedges, L., & Citkowicz, V. (2015). Estimating effect size when there is clustering in one treatment group. *Behavior Research Methods, 47*(4), 1295-1308. doi: 10.3758/s13428-014-0538-z

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60-87. doi:10.3102/0162373707299706

Ho, D., Imai, K., King, G., & Stuart, E. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15*(3), 199-236. doi: 10.1093/pan/mpl013

Hogg, R. (1974). Adaptive robust procedures: A partial review and some suggestions for future applications and theory. *Journal of the American Statistical Association, 69*(348), 909-923. doi**:** 10.1080/01621459.1974.10480225

Hox, J. (2010). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Jason, L., & Glenwick, D. (2016). *Handbook of methodological approaches to community-based research: Qualitative, quantitative, and mixed methods*. New York, NY: Oxford University Press.

Ketelsen, K. L. (2014). *A monte carlo simulation to examine the effects of violating the normality assumption in 2-level hierarchical linear models with unbalanced designs* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (1625404890)

Kim, K. (1990). *Multilevel data analysis: A comparative examination of analytical alternatives* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (303889116)

Kloke, J., & McKean, J. (2014). npsm: Package for Nonparametric Statistical Methods using R (R package version 3, 1)[Software]. Available from https://cran.r-project.org/web/packages/npsm/npsm.pdf

Konstantopoulos, S. (2009). Incorporating cost in power analysis for three-level cluster-randomized designs. *Evaluation Review, 33*(4), 335-357. doi: 10.1177/0193841X09337991

Korendijk, E.J.H., Maas, C.J.M., Hox, J. & Moerbeek, M. (2012). The robustness of the parameter and standard error estimates in trials with partially nested data: A simulation study. In E. Korendijk (Ed.). *Robustness and optimal design issues for cluster randomized trials* (pp. 59-94). Dissertation, Utrecht University. Retrieved from https://dspace.library.uu.nl/handle/1874/240965

Kreft I. G. G. (1996) Are multilevel techniques necessary? An overview, including 19 simulation studies. Unpublished manuscript, California State University at Los Angeles.

Krull, J. L. (1997). *The effects of misspecification resulting from analysis decisions in multilevel models* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (304339280)

Kuznetsova, A., Brockhoff, P.B. and Christensen, R.H.B. (2015). Package 'lmerTest'. (version 2) [R package] Available from https://cran.opencpu.org/web/packages/lmerTest/lmerTest.pdf

LeBeau, B. (2013). *Misspecification of the covariance matrix in the linear mixed model: A monte carlo simulation* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (1322973438)

Lee, Y., Nelder, J.A. & Pawitan, Y. (2006). *Generalized linear models with random effects:Unified analysis via H-likelihood.* Boca Raton: Chapman & Hall/CRC.

Lee, K. J., & Thompson, S. G. (2005). The use of random effects models to allow for clustering in individually randomized trials. *Clinical Trials*, *2*(2), 163-173. Retrieved from http://journals.sagepub.com.ezp1.lib.umn.edu/doi/pdf/10.1191/1740774505cn082oa

Lesaux, N. K., Kieffer, M. J., Kelley, J. G., & Harris, J. R. (2014). Effects of academic vocabulary instruction for linguistically diverse adolescents evidence from a randomized field trial. *American Educational Research Journal*, *51*(6), 1159-1194. doi: 0002831214532165.

Lohr, S. (2009). *Sampling: Design and analysis*. Boston, MA: Brooks/Cole.

Lohr, S., Schochet, P. Z., & Sanders, E. (2014). *Partially nested randomized controlled trials in education research: A guide to design and analysis* (No. 7dce2a40502a473eb80425d5ea970ae3). Mathematica Policy Research. Retrieved from http://ies.ed.gov/ncer/pubs/

Luo, W., Cappaert, K., & Ning, L. (2015). Modelling partially cross-classified multilevel data. *British Journal of Mathematical and Statistical Psychology*, *68*(2), 342-362. doi:10.1111/bmsp.12050

Maas, C. J., & Hox, J. J. (2004a). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis*, *46*(3), 427-440. doi : 10.1016/j.csda.2003.08.006

Maas, C.J.M. & Hox, J.J. (2004b). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, *58*(2), 127-137. doi:10.1046/j.0039-0402.2003.00252.x

Maeda, Y. (2007). *Monte carlo evidence regarding the effects of violating assumed conditions of two-level hierarchical models for cross -sectional data* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses. (304840748)

Maravina, T. (2012). *Tests for differences between least squares and robust regression parameter estimates and related topics* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses. (3552826)

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156-166. doi : 10.1037/0033-2909.105.1.156

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research*, *39*(1), 129-149. doi:10.1207/s15327906mbr3901_5

Morgan, S., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. New York, NY: Cambridge University Press.

Orr, L. L. (1999). *Social experiments: Evaluating public programs with experimental methods*. Thousand Oaks, CA.: Sage Publications.

Osborne, J. W. (2000). Advantages of hierarchical linear modeling. *Practical Assessment, Research & Evaluation*, *7*(1), 1-3. Retrieved from http://pareonline.net/getvn.asp?v=7&n=1

Paccagnella, O. (2011). Sample size and accuracy of estimates in multilevel models: New simulation results. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 7*(3), 111-120. doi: 10.1027/1614-2241/a000029

Pearson, E.S. (1931). The analysis of variance in cases of non-normal variation. *Biometrika*, 23, 114-133. doi: 10.2307/2333631

Peña, D., Zamar, R., & Yan, G. (2009). Bayesian likelihood robustness in linear models. *Journal of Statistical Planning and Inference, 139*(7), 2196-2207. doi : 10.1016/j.jspi.2008.10.012

Posner, J. K., & Vandell, D. L. (1994). Low-income children's after-school care: Are there beneficial effects of after-school programs? *Child Development*, 65, 440–456. doi:10.1111/j.1467-8624.1994.tb00762.x

R Core Team. (2015). R: A language and environment for statistical computing (Version 3.2.2) [Computer software manual]. Viena, Austria. Retrieved from http://www.R-project.org/

Raab, G. M., & Butcher, I. (2001). Balance in cluster randomized trials. *Statistics in Medicine*, *20*(3), 351-365. doi**:**10.1002/1097-0258(20010215)20:3<351::AID-SIM797>3.0.CO;2-C

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd Ed.). Newbury Park, CA: Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R., & du Toit, M. (2011). HLM statistical software: (Version 7). [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

Roberts, C. (1999). The implications of variation in outcome between health professionals for the design and analysis of randomized controlled trials. *Statistics in Medicine, 18*(19), 2605-2615. doi**:**10.1002/(SICI)1097-0258(19991015)18:19<2605::AID-SIM237>3.0.CO;2-N

Roberts, C., & Roberts, S. A. (2005). Design and analysis of clinical trials with clustering effects due to treatment. *Clinical Trials*, *2*(2), 152-162. Retrieved from http://journals.sagepub.com.ezp1.lib.umn.edu/doi/pdf/10.1191/1740774505cn076oa

Roberts, J., Williams, K., Carter, M., Evans, D., Parmenter, T., Silove, N., Clark, T., & Warren, A. (2011). A randomized controlled trial of two early intervention programs for young children with autism: Centre-based with parent program and home-based. *Research in Autism Spectrum Disorders*, *5*(4), 1553-1566.doi: 10.1016/j.rasd.2011.03.001

Roberts, L. W., Geppert, C., Connor, R., Nguyen, K., & Warner, T. D. (2001). An invitation for medical educators to focus on ethical and policy issues in research and scholarly practice. *Academic Medicine*, *76*(9), 876-885. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.470.3603&rep=rep1&type=pdf

Roberts, L. W., Geppert, C. M., Coverdale, J., Louie, A., & Edenharder, K. (2005). Ethical and regulatory considerations in educational research. *Academic Psychiatry*, *29*(1), 1-5.doi: 10.1176/appi.ap.29.1.1

Robinson, L., & Jewell, N. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review / Revue Internationale De Statistique, 59*(2), 227-240. doi: 10.2307/1403444

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688. doi: 10.1037/h0037350

Savage, R. S., Abrami, P., Hipps, G., & Deault, L. (2009). A randomized controlled trial study of the ABRACADABRA reading intervention program in grade 1. *Journal of Educational Psychology*, *101*(3), 590. doi: 10.1037/a0014700

Sanders, E. A. (2011). *Multilevel analysis methods for partially nested cluster randomized trials* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (865641994)

Schochet, Peter Z. (2008). Statistical Power for Random Assignment Evaluations of Education Programs. *Journal of Educational and Behavioral Statistics, 33*(1), 62-87. doi: 10.3102/1076998607302714

Seltzer, M. (1993). Sensitivity analysis for fixed effects in the hierarchical model: A Gibbs sampling approach. *Journal of Educational and Behavioral Statistics*, *18*(3), 207-235. doi : http://dx.doi.org/10.3102/10769986018003207

Serlin, R. C., Wampold, B. E., & Levin, J. R. (2003). Should providers of treatment be regarded as a random factor? If it ain't broke, don't "fix" it: A comment on Siemer and Joorman (2003). *Psychological Methods, 8,* 524–534. doi:10.1037/1082-989X.8.4.524

Shieh, Y. (1999). *An evaluation of mixed effects multilevel modeling under conditions of error term nonnormality* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (9947388)

Shieh, Y.Y., Fouladi, R.T., & Pullum, T.W. (2001). The effect of error term non-normality on multilevel model parameter estimates and standard errors: A focus on estimation bias. *Multiple Linear Regression Viewpoints*, *27*(1), 12-37.

Shieh, Y.Y. & Fouladi, R.T. (2003). The effect of multicollinearity on multilevel modeling parameter estimates and standard errors. *Educational and Psychological Measurement*, 63(6), 951-985. Retrieved from http://journals.sagepub.com.ezp1.lib.umn.edu/doi/pdf/10.1177/0013164403258402

Shih, T. (2008). *Adequate sample sizes for viable 2-level hierarchical linear modeling analysis: A study on sample size requirement in HLM in relation to different intraclass correlations* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (304435724)

Siemer, M., & Joorman, J. (2003a). Assumptions and consequences of treating providers in therapy studies as fixed versus random effects: Reply to Crits-Christoph, Tu, and Gallop (2003) and Serline, Wampold, and Levin (2003). *Psychological Methods, 8,* 535–544. doi:10.1037/ 1082-989X.8.4.535

Siemer, M., & Joorman, J. (2003b). Power and measures of effect size in analysis of variance with fixed versus random nested factors. *Psychological Methods, 8,* 497–517. doi:10.1037/1082-989X.8.4.524s

Singer, J. (1987). An intraclass correlation model for analyzing multilevel data. *The Journal of Experimental Education*, *55*(4), 219-228. doi; 10.1080/00220973.1987.10806457

Snijders, T.A.B. (2005). Power and sample size in multilevel linear models. In: B.S. Everitt and Howell (eds), *Encyclopedia of statistics in behavioral science*. Volume 3, 1570-1573. Chicester (etc). Willey, 2005.

Snijders, T.A.B., & Bosker, R.J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics*, 18, 237-259. Retrieved from https://www.stats.ox.ac.uk/~snijders/SnijdersBosker1993.pdf.

Snijders, T.A.B., & Bosker, R.J. (1999) *Multilevel analysis: An introduction to basic and advance multilevel modeling.* Thousand Oaks, CA: Sage Publications.

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., Raudenbush, S., & TO, A. (2011). *Optimal design plus empirical evidence: Documentation for the "Optimal Design" software.* William T. Grant Foundation. Retrieved http://hlmsoft.net/od/od-manual-20111016-v300.pdf.

Talley A.E. (2013). *The impact of nonnormal and heteroscedastic level one residuals in partially clustered data.* Retrieved from https://repositories.lib.utexas.edu/handle/2152/22630

Tessler, J. M. (2014). *Three-level models for partially nested data structures*. Retrieved from https://escholarship.org/uc/item/21q2m89j

Trochim, W. M. (2005). *Research methods: The concise knowledge base*. Mason, OH: Thomson Custom Pub.

Van der Leeden, R., & Busing, F. (1994*). First iteration versus IGLS RIGLS estimates in two-level models: A Monte Carlo study with ML3*. Unpublished manuscript, Leiden University, the Netherlands.

Van der Leeden, R., Busing, F., & Meijer, E. (1997, April). *Applications of bootstrap methods for two-level models*. Paper presented at the Multilevel Conference, Amsterdam.

Verbeke, G. & Lesaffre, E. (1997). The effect of misspecifying the random effects distribution in linear mixed models for longitudinal data. *Computational Statistics and Data Analysis*, 23, 541-556. doi: 10.1016/S0167-9473(96)00047-3

Wieand, G. & Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika, 71*(3), 431-444. doi: 10.2307/2336553

West, B., Welch, K., & Gałecki, A. (2007). *Linear mixed models: A practical guide using statistical software*. Boca Raton: Chapman & Hall/CRC.

Zar, J. (1996). *Biostatistical analysis* (3rd ed.). Upper Saddle River, N.J.: Prentice-Hall.

Zepeda, C. D., Richey, J. E., Ronevich, P., & Nokes-Malach, T. J. (2015). Direct instruction of metacognition benefits adolescent science learning, transfer, and motivation: An in vivo study. Abstract retrieved from http://dx.doi.org/10.1037/edu0000022

Zhang, D. (2005). *A Monte Carlo investigation of robustness to nonnormal incomplete data of multilevel modeling* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (3231609)

Zopluoğlu, C. (2012). A cross-national comparison of intra-class correlation coefficient in educational achievement outcomes. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *3*(5). Retrieved from http://www.epod-online.org/sayilar/sayi5/makale4.pdf

Appendix A

This appendix contains tables for the ANOVA analysis performed for the relative bias of the fixed effects, power and Type I error rated, and for the WLS regression analysis performed on ln(RMSE).

Table 28

ANOVA for the γ_{10} absolute and relative bias in Experiment 1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Distributions	2	0.0002	1.00E-04	0.6060	0.5580
Clusters	2	0.0003	2.00E-04	0.8010	0.4660
Cluster size	2	0.0005	3.00E-04	1.2240	0.3200
ICC	2	0.0002	1.00E-04	0.3940	0.6810
Distributions × Clusters	4	0.0018	5.00E-04	2.1980	0.1160
Distributions × Cluster size	4	0.0004	1.00E-04	0.4720	0.7560
Distributions × ICC	4	0.0002	1.00E-04	0.2980	0.8750
Clusters × Cluster size	4	0.0005	1.00E-04	0.5720	0.6870
Clusters × ICC	4	0.0001	0.00E+00	0.1720	0.9490
Cluster size × ICC	4	0.0015	4.00E-04	1.8090	0.1760
Distributions× Clusters × Cluster size	8	0.0019	2.00E-04	1.1490	0.3850
Distributions× Clusters × ICC	8	0.0003	0.00E+00	0.1760	0.9910
Distributions × Cluster size × ICC	8	0.0008	1.00E-04	0.4890	0.8460
Clusters × Cluster size × ICC	8	0.0016	2.00E-04	0.9880	0.4800
Residuals	16	0.0033	2.00E-04		

Note: Significance codes are 0 '***', 0.001 '**', 0.01 '*', 0.05 '.' , 0.1 ' ' 1

Table 29

Weighted Least Square Regression for the γ_{10} ln(RMSE) in Experiment 1

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.1880	0.1720	1.0930	0.2790	
$t_{(df = 11)}$	0.0865	0.2160	0.4010	0.6902	
$t_{(df = 4)}$	0.4150	0.2160	1.9200	0.0597	
Cluster	-0.0201	0.0047	-4.2770	0.0001	***
Cluster size	-0.0252	0.0075	-3.3550	0.0014	**
ICC	-3.6400	0.9420	-3.8670	0.0003	***
$t_{(df = 11)}$ × Clusters	0.0003	0.0055	0.0600	0.9524	
$t_{(df = 4)}$ × Clusters	-0.0022	0.0055	-0.3890	0.6985	
$t_{(df = 11)}$ × Cluster size	0.0006	0.0087	0.0670	0.9466	
$t_{(df = 4)}$ × Cluster size	-0.0015	0.0087	-0.1740	0.8625	
$t_{(df = 11)}$ × ICC	0.2950	1.1100	0.2670	0.7905	
$t_{(df = 4)}$ × ICC	0.0709	1.1100	0.0640	0.9491	
Clusters × Cluster size	0.0000	0.0002	0.0770	0.9391	
Clusters × ICC	0.0042	0.0249	0.1690	0.8660	
Cluster size × ICC	0.0710	0.0395	1.7980	0.0772	
$t_{(df = 11)}$× Clusters × Cluster size	0.0000	0.0002	0.0150	0.9883	
$t_{(df = 4)}$× Clusters × Cluster size	0.0001	0.0002	0.2900	0.7727	
$t_{(df = 11)}$× Clusters × ICC	-0.0021	0.0250	-0.0850	0.9324	
$t_{(df = 4)}$× Clusters × ICC	0.0021	0.0250	0.0830	0.9345	
$t_{(df = 11)}$× Cluster size × ICC	-0.0115	0.0383	-0.3010	0.7647	
$t_{(df = 4)}$× Cluster size × ICC	-0.0067	0.0383	-0.1760	0.8611	
Clusters × Cluster size × ICC	-0.0002	0.0010	-0.1870	0.8524	

Note: Significance codes are 0 '***', 0.001 '**', 0.01 '*', 0.05 '.' , 0.1 ' ' 1

Table 30

ANOVA for the power of γ_{10} in Experiment 1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Distributions	2	0.3350	1.68E-01	239.5380	0.0000	***
Clusters	2	3.3440	1.67E+00	2390.4400	0.0000	***
Cluster size	2	0.6200	3.10E-01	443.2770	0.0000	***
ICC	2	0.8240	4.12E-01	589.1190	0.0000	***
Distributions × Clusters	4	0.0230	5.70E-03	8.1490	0.0009	***
Distributions × Cluster size	4	0.0070	1.90E-03	2.6550	0.0714	.
Distributions × ICC	4	0.0120	3.10E-03	4.3740	0.0140	*
Clusters × Cluster size	4	0.0250	6.20E-03	8.8950	0.0006	***
Clusters × ICC	4	0.0420	1.04E-02	14.8470	0.0000	***
Cluster size × ICC	4	0.1620	4.06E-02	58.0550	0.0000	***
Distributions× Clusters × Cluster size	8	0.0160	2.10E-03	2.9380	0.0317	*
Distributions× Clusters × ICC	8	0.0230	2.90E-03	4.0880	0.0080	**
Distributions × Cluster size × ICC	8	0.0030	4.00E-04	0.5030	0.8369	
Clusters × Cluster size × ICC	8	0.0620	7.80E-03	11.1270	0.0000	***
Residuals	16	0.0110	7.00E-04			

Note: Significance codes are 0 ‘***’, 0.001 ‘**’, 0.01 ‘*’, 0.05 ‘.’ , 0.1 ‘ ’ 1

Table 31

Simplified ANOVA for the power of γ_{10} in Experiment 1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	η_p^2	
Distributions	2	0.3350	1.68E-01	287.1290	0.0001	0.0610	.
Clusters	2	3.3440	1.67E+00	2865.3700	0.0001	0.6070	***
Cluster size	2	0.6200	3.10E-01	531.3460	0.0001	0.1130	*
ICC	2	0.8240	4.12E-01	706.1640	0.0001	0.1500	**
Distributions × Clusters	4	0.0230	5.70E-03	9.7680	0.0001	0.0040	.
Distributions × ICC	4	0.0120	3.10E-03	5.2420	0.0035	0.0020	.
Clusters × Cluster size	4	0.0250	6.20E-03	10.6620	0.0000	0.0050	.
Clusters × ICC	4	0.0420	1.04E-02	17.7970	0.0000	0.0080	.
Cluster size × ICC	4	0.1620	4.06E-02	69.5890	0.0000	0.0290	.
Distributions× Clusters × Cluster size	12	0.0240	2.00E-03	3.4080	0.0051	0.0040	.
Distributions× Clusters × ICC	8	0.0230	2.90E-03	4.9010	0.0011	0.0040	.
Clusters × Cluster size × ICC	8	0.0620	7.80E-03	13.3380	0.0000	0.0110	.
Residuals	24	0.0140	6.00E-04				

Note: . negligible effect, * small effect, ** medium effect, and *** large effect

Table 32

ANOVA for the type I error rate of γ_{10} in Experiment 1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Distributions	2	0.0002	8.66E-05	1.8990	0.1820	
Clusters	2	0.0000	3.20E-06	0.0700	0.9325	
Cluster size	2	0.0004	2.03E-04	4.4490	0.0291	*
ICC	2	0.0002	8.88E-05	1.9470	0.1751	
Distributions × Clusters	4	0.0003	6.70E-05	1.4700	0.2576	
Distributions × Cluster size	4	0.0004	8.84E-05	1.9400	0.1528	
Distributions × ICC	4	0.0005	1.17E-04	2.5570	0.0790	
Clusters × Cluster size	4	0.0002	4.77E-05	1.0470	0.4142	
Clusters × ICC	4	0.0003	8.60E-05	1.8860	0.1621	
Cluster size × ICC	4	0.0002	4.60E-05	1.0090	0.4319	
Distributions× Clusters × Cluster size	8	0.0006	6.92E-05	1.5190	0.2267	
Distributions× Clusters × ICC	8	0.0003	3.10E-05	0.6810	0.7023	
Distributions × Cluster size × ICC	8	0.0004	4.64E-05	1.0180	0.4610	
Clusters × Cluster size × ICC	8	0.0001	9.49E-06	0.2080	0.9848	
Residuals	16	0.0007	4.56E-05			

Note: Significance codes are 0 '***', 0.001 '**', 0.01 '*', 0.05 '.' , 0.1 ' ' 1

Table 33

Simplified ANOVA for the type I error rate of γ_{10} in Experiment 1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	η_p^2	
Cluster size	2	0.0004	2.03E-04	3.8190	0.0262	0.0890	.
Residuals	78	0.0041	5.31E-05				

Note: . negligible effect, * small effect, ** medium effect, and *** large effect

Table 34

ANOVA for the γ_{10} relative bias in Experiment 2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Distributions	2	0.0009	5.00E-04	1.7550	0.2046
Clusters	2	0.0014	7.00E-04	2.7170	0.0964 .
Cluster size	2	0.0001	0.00E+00	0.1140	0.8931
ICC	2	0.0002	1.00E-04	0.3800	0.6901
Distributions × Clusters	4	0.0011	3.00E-04	1.0520	0.4118
Distributions × Cluster size	4	0.0005	1.00E-04	0.5040	0.7331
Distributions × ICC	4	0.0004	1.00E-04	0.3860	0.8157
Clusters × Cluster size	4	0.0003	1.00E-04	0.2600	0.8995
Clusters × ICC	4	0.0031	8.00E-04	2.8920	0.0561 .
Cluster size × ICC	4	0.0003	1.00E-04	0.2910	0.8798
Distributions× Clusters × Cluster size	8	0.0023	3.00E-04	1.0780	0.4248
Distributions× Clusters × ICC	8	0.0019	2.00E-04	0.9110	0.5314
Distributions × Cluster size × ICC	8	0.0009	1.00E-04	0.4460	0.8755
Clusters × Cluster size × ICC	8	0.0012	2.00E-04	0.5820	0.7782
Residuals	16	0.0042	3.00E-04		

Note: Significance codes are 0 '***', 0.001 '**', 0.01 '*', 0.05 '.' , 0.1 ' ' 1

Table 35

ANOVA for the γ_{20} relative bias in Experiment 2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Distributions	2	0.0000	5.95E-06	0.321	0.7299	
Clusters	2	0.0002	7.77E-05	4.193	0.0344	*
Cluster size	2	0.0002	1.22E-04	6.582	0.0082	**
ICC	2	0.0002	9.94E-05	5.36	0.0165	*
Distributions × Clusters	4	0.0000	1.08E-05	0.582	0.6802	
Distributions × Cluster size	4	0.0000	5.63E-06	0.304	0.8713	
Distributions × ICC	4	0.0000	5.80E-06	0.313	0.8651	
Clusters × Cluster size	4	0.0003	6.23E-05	3.361	0.0354	*
Clusters × ICC	4	0.0003	6.60E-05	3.561	0.0293	*
Cluster size × ICC	4	0.0004	1.09E-04	5.883	0.0042	**
Distributions× Clusters × Cluster size	8	0.0003	3.86E-05	2.084	0.1005	
Distributions× Clusters × ICC	8	0.0000	4.79E-06	0.258	0.9709	
Distributions × Cluster size × ICC	8	0.0002	3.03E-05	1.634	0.1918	
Clusters × Cluster size × ICC	8	0.0010	1.29E-04	6.965	0.0005	***
Residuals	16	0.0003	1.85E-05			

Note: Significance codes are 0 '***', 0.001 '**', 0.01 '*', 0.05 '.' , 0.1 ' ' 1

Table 36

Simplified ANOVA for the γ_{20} relative bias in Experiment 2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	η_p^2	
Clusters	2	0.0002	7.77E-05	4.252	0.01928	0.044	.
Cluster size	2	0.0002	1.22E-04	6.675	0.00257	0.068	.
ICC	2	0.0002	9.94E-05	5.435	0.00707	0.056	.
Clusters × Cluster size	4	0.0003	6.23E-05	3.408	0.01477	0.07	.
Clusters × ICC	4	0.0003	6.60E-05	3.611	0.01113	0.074	.
Cluster size × ICC	4	0.0004	1.09E-04	5.966	0.00047	0.122	*
Clusters × Cluster size × ICC	8	0.0010	1.29E-04	7.063	2.42E-06	0.29	***
Residuals	54	0.0010	1.83E-05				

Note: . negligible effect, * small effect, ** medium effect, and *** large effect

Table 37

Weighted Least Square Regression for the γ_{10} ln(RMSE) in Experiment 2

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.1570	0.1700	0.926	0.3580	
$t_{(df = 11)}$	0.1510	0.2130	0.708	0.4819	
$t_{(df = 4)}$	0.3860	0.2130	1.811	0.0752	
Cluster	-0.0194	0.0046	-4.171	0.0001	***
Cluster size	-0.0233	0.0074	-3.148	0.0026	**
ICC	-3.3200	0.9280	-3.576	0.0007	***
$t_{(df = 11)}$ × Clusters	-0.0010	0.0054	-0.179	0.8587	
$t_{(df = 4)}$ × Clusters	-0.0006	0.0054	-0.117	0.9070	
$t_{(df = 11)}$ × Cluster size	-0.0024	0.0086	-0.275	0.7847	
$t_{(df = 4)}$ × Cluster size	-0.0016	0.0086	-0.191	0.8495	
$t_{(df = 11)}$ × ICC	-0.1080	1.0900	-0.099	0.9216	
$t_{(df = 4)}$ × ICC	-0.1930	1.0900	-0.177	0.8598	
Clusters × Cluster size	0.0000	0.0002	-0.22	0.8270	
Clusters × ICC	-0.0028	0.0245	-0.115	0.9090	
Cluster size × ICC	0.0617	0.0389	1.587	0.1179	
$t_{(df = 11)}$× Clusters × Cluster size	0.0000	0.0002	0.232	0.8173	
$t_{(df = 4)}$× Clusters × Cluster size	0.0000	0.0002	0.177	0.8602	
$t_{(df = 11)}$× Clusters × ICC	0.0015	0.0246	0.061	0.9514	
$t_{(df = 4)}$× Clusters × ICC	0.0012	0.0246	0.049	0.9613	
$t_{(df = 11)}$× Cluster size × ICC	0.0013	0.0377	0.034	0.9727	
$t_{(df = 4)}$× Cluster size × ICC	0.0046	0.0377	0.123	0.9025	
Clusters × Cluster size × ICC	0.0001	0.0009	0.066	0.9475	

Note: Significance codes are 0 '***', 0.001 '**', 0.01 '*', 0.05 '.' , 0.1 ' ' 1

Table 38

Weighted Least Square Regression for the γ_{20} ln(RMSE) in Experiment 2

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.3700	0.2120	-1.7490	0.0855	.
$t_{(df = 11)}$	0.0310	0.2660	0.1170	0.9076	
$t_{(df = 4)}$	0.3160	0.2660	1.1900	0.2389	
Cluster	-0.0223	0.0058	-3.8460	0.0003	***
Cluster size	-0.0361	0.0092	-3.9020	0.0002	***
ICC	-4.7400	1.1600	-4.0930	0.0001	***
$t_{(df = 11)}$ × Clusters	0.0018	0.0068	0.2590	0.7965	
$t_{(df = 4)}$ × Clusters	0.0011	0.0068	0.1650	0.8694	
$t_{(df = 11)}$ × Cluster size	0.0024	0.0107	0.2270	0.8211	
$t_{(df = 4)}$ × Cluster size	0.0009	0.0107	0.0860	0.9318	
$t_{(df = 11)}$ × ICC	0.2100	1.3600	0.1550	0.8775	
$t_{(df = 4)}$ × ICC	0.0558	1.3600	0.0410	0.9674	
Clusters × Cluster size	0.0001	0.0002	0.4680	0.6414	
Clusters × ICC	0.0060	0.0306	0.1970	0.8444	
Cluster size × ICC	0.0091	0.0486	0.1880	0.8514	
$t_{(df = 11)}$× Clusters × Cluster size	-0.0001	0.0002	-0.2750	0.7844	
$t_{(df = 4)}$× Clusters × Cluster size	0.0000	0.0002	-0.1560	0.8767	
$t_{(df = 11)}$× Clusters × ICC	-0.0033	0.0307	-0.1080	0.9143	
$t_{(df = 4)}$× Clusters × ICC	-0.0022	0.0307	-0.0700	0.9445	
$t_{(df = 11)}$× Cluster size × ICC	-0.0018	0.0471	-0.0380	0.9699	
$t_{(df = 4)}$× Cluster size × ICC	0.0029	0.0471	0.0610	0.9517	
Clusters × Cluster size × ICC	-0.0003	0.0012	-0.2600	0.7956	

Note: Significance codes are 0 '***', 0.001 '**', 0.01 '*', 0.05 '.' , 0.1 ' ' 1

Table 39

ANOVA for the power of γ_{10} in Experiment 2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Distributions	2	0.3410	1.71E-01	311.0110	0.000	***
Clusters	2	3.5120	1.76E+00	3202.5200	0.000	***
Cluster size	2	0.6270	3.14E-01	572.0210	0.000	***
ICC	2	0.7790	3.90E-01	710.6600	0.000	***
Distributions × Clusters	4	0.0290	7.20E-03	13.0590	0.000	***
Distributions × Cluster size	4	0.0080	2.00E-03	3.5890	0.028	*
Distributions × ICC	4	0.0090	2.20E-03	3.9600	0.020	*
Clusters × Cluster size	4	0.0260	6.60E-03	12.0200	0.000	***
Clusters × ICC	4	0.0300	7.40E-03	13.5370	0.000	***
Cluster size × ICC	4	0.1730	4.32E-02	78.8480	0.000	***
Distributions× Clusters × Cluster size	8	0.0170	2.10E-03	3.8220	0.011	*
Distributions× Clusters × ICC	8	0.0210	2.60E-03	4.7180	0.004	**
Distributions × Cluster size × ICC	8	0.0040	5.00E-04	0.9400	0.512	
Clusters × Cluster size × ICC	8	0.0600	7.40E-03	13.5690	0.000	***
Residuals	16	0.0090	5.00E-04			

Note: Significance codes are 0 ‘***’, 0.001 ‘**’, 0.01 ‘*’, 0.05 ‘.’ , 0.1 ‘ ’ 1

Table 40

Simplified ANOVA for the power of γ_{10} in Experiment 2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	η_p^2	
Distributions	2	0.3410	1.71E-01	317.328	0.0000	0.0600	.
Clusters	2	3.5120	1.76E+00	3267.570	0.0000	0.6220	***
Cluster size	2	0.6270	3.14E-01	583.640	0.0000	0.1110	*
ICC	2	0.7790	3.90E-01	725.095	0.0000	0.1380	*
Distributions × Clusters	4	0.0290	7.20E-03	13.324	0.0000	0.0050	.
Distributions × Cluster size	4	0.0080	2.00E-03	3.662	0.0182	0.0010	.
Distributions × ICC	4	0.0090	2.20E-03	4.040	0.0121	0.0020	.
Clusters × Cluster size	4	0.0260	6.60E-03	12.264	0.0000	0.0050	.
Clusters × ICC	4	0.0300	7.40E-03	13.812	0.0000	0.0050	.
Cluster size × ICC	4	0.1730	4.32E-02	80.449	0.0000	0.0310	.
Distributions× Clusters × Cluster size	8	0.0170	2.10E-03	3.900	0.0045	0.0030	.
Distributions× Clusters × ICC	8	0.0210	2.60E-03	4.814	0.0013	0.0040	.
Clusters × Cluster size × ICC	8	0.0600	7.40E-03	13.845	0.0000	0.0110	.
Residuals	24	0.0130	5.00E-04				

Note: . negligible effect, * small effect, ** medium effect, and *** large effect

Table 41

ANOVA for the power of γ_{20} in Experiment 2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Distributions	2	0.0110	5.52E-03	12.4400	0.0006	***
Clusters	2	0.0698	3.49E-02	78.6840	0.0000	***
Cluster size	2	0.0676	3.38E-02	76.1950	0.0000	***
ICC	2	0.0749	3.74E-02	84.4090	0.0000	***
Distributions × Clusters	4	0.0103	2.57E-03	5.7940	0.0044	**
Distributions × Cluster size	4	0.0089	2.23E-03	5.0360	0.0081	**
Distributions × ICC	4	0.0119	2.98E-03	6.7200	0.0023	**
Clusters × Cluster size	4	0.0876	2.19E-02	49.3660	0.0000	***
Clusters × ICC	4	0.0991	2.48E-02	55.8830	0.0000	***
Cluster size × ICC	4	0.0955	2.39E-02	53.8730	0.0000	***
Distributions× Clusters × Cluster size	8	0.0058	7.20E-04	1.6350	0.1916	
Distributions× Clusters × ICC	8	0.0084	1.06E-03	2.3800	0.0666	.
Distributions × Cluster size × ICC	8	0.0074	9.30E-04	2.0950	0.0991	.
Clusters × Cluster size × ICC	8	0.1139	1.42E-02	32.1070	0.0000	***
Residuals	16	0.0071	4.40E-04			

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 42: *Simplified ANOVA for the power of γ_{20} in Experiment 2*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	η_p^2	
Distributions	2	0.0110	5.52E-03	7.67	0.0015	0.0162	.
Clusters	2	0.0698	3.49E-02	48.514	0.0000	0.1027	*
Cluster size	2	0.0676	3.38E-02	46.98	0.0000	0.0995	*
ICC	2	0.0749	3.74E-02	52.044	0.0000	0.1102	*
Distributions × Clusters	4	0.0103	2.57E-03	3.572	0.0139	0.0151	.
Distributions × Cluster size	4	0.0089	2.23E-03	3.105	0.0257	0.0132	.
Distributions × ICC	4	0.0119	2.98E-03	4.143	0.0067	0.0176	.
Clusters × Cluster size	4	0.0876	2.19E-02	30.438	0.0000	0.1289	*
Clusters × ICC	4	0.0991	2.48E-02	34.456	0.0000	0.1459	**
Cluster size × ICC	4	0.0955	2.39E-02	33.217	0.0000	0.1407	**
Clusters × Cluster size × ICC	8	0.1139	1.42E-02	19.796	0.0000	0.1677	**
Residuals	40	0.0288	7.20E-04				

Note: . negligible effect, * small effect, ** medium effect, and *** large effect

Table 43

ANOVA of the γ_{10} Type I error rate by conditions, in Experiment 2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Distributions	2	3.40E-05	1.69E-05	0.447	0.647	
Clusters	2	0.00012	6.04E-05	1.596	0.233	
Cluster size	2	0.00018	8.94E-05	2.364	0.126	
ICC	2	0.00027	1.36E-04	3.581	0.052	.
Distributions × Clusters	4	0.00013	3.21E-05	0.848	0.516	
Distributions × Cluster size	4	0.00073	1.84E-04	4.856	0.009	**
Distributions × ICC	4	8.70E-05	2.16E-05	0.572	0.687	
Clusters × Cluster size	4	0.00046	1.15E-04	3.046	0.048	*
Clusters × ICC	4	7.40E-05	1.85E-05	0.488	0.745	
Cluster size × ICC	4	0.00037	9.23E-05	2.44	0.089	.
Distributions× Clusters × Cluster size	8	0.00038	4.74E-05	1.252	0.333	
Distributions× Clusters × ICC	8	0.00022	2.72E-05	0.719	0.673	
Distributions × Cluster size × ICC	8	0.00022	2.77E-05	0.732	0.663	
Clusters × Cluster size × ICC	8	0.00024	2.99E-05	0.79	0.619	
Residuals	16	0.00061	3.78E-05			

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 44

Simplified ANOVA of the γ_{10} Type I error rate by conditions, in Experiment 2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)		η_p^2
Distributions × Cluster size	8	0.0010	1.18E-04	3.0170	0.0060	***	0.2300
Clusters × Cluster size	6	0.0006	9.70E-05	2.470	0.0324	**	0.1410
Residuals	66	0.0026	3.93E-05				

Note: . negligible effect, * small effect, ** medium effect, and *** large effect

Table 45

ANOVA of the γ_{20} Type I error rate by conditions, in Experiment 2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Distributions	2	0.0004	1.74E-04	5.421	0.0159 *
Clusters	2	0.0002	7.83E-05	2.44	0.1189
Cluster size	2	0.0002	9.61E-05	2.994	0.0786 .
ICC	2	0.0001	5.87E-05	1.828	0.1927
Distributions × Clusters	4	0.0001	2.26E-05	0.705	0.6001
Distributions × Cluster size	4	0.0000	5.07E-06	0.158	0.9565
Distributions × ICC	4	0.0001	1.39E-05	0.434	0.7819
Clusters × Cluster size	4	0.0001	1.64E-05	0.509	0.7298
Clusters × ICC	4	0.0001	1.75E-05	0.545	0.7052
Cluster size × ICC	4	0.0000	1.02E-05	0.317	0.8627
Distributions× Clusters × Cluster size	8	0.0004	4.75E-05	1.48	0.2398
Distributions× Clusters × ICC	8	0.0004	4.86E-05	1.514	0.2283
Distributions × Cluster size × ICC	8	0.0002	2.43E-05	0.758	0.6429
Clusters × Cluster size × ICC	8	0.0003	3.38E-05	1.054	0.439
Residuals	16	0.0005	3.21E-05		

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 46

Simplified ANOVA of the γ_{20} Type I error rate by conditions, in Experiment

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	η_p^2
Distributions	2	0.000	1.74E-04	5.309	.007	0.12*
Residuals	78	0.003	3.28E-05			

Note: . negligible effect, * small effect, ** medium effect, and *** large effect

Appendix B

This appendix contains the simulation codes to generate data for models A and for model B in order to estimate the average Type I error rate, Power, and fixed effects.

Simulation codes for A model: Type I error rate

```
############################################################
##### Codes for simulating data NORMAL DISTRIBUTION ########

##### invoke the packages you need

library(lme4) # Fit the model
library(lmerTest) # estimate the p-values
library(dplyr)

## (1) define your conditions

x<-c(10,30, 50) ### level-2 sample size
y<-c(6, 17, 32) ### within cluster sample size
z<-c(.05,.15,.25) ### intra-class correlation treatment condition
d<-c("Normal", "t4", "t11")


## (2) design your matrix of conditions

condi<- expand.grid(x,y,z,d)
print(condi)

## (3) write your function (inner function): in my case it should be the
## (a) simulation of clusters
## (b) simulation of subjects
## (c) simulation or error distribution
## (d) etc.
set.seed(011179)

simul <-function(x, y, z, d){


 if(d=="Normal"){
 data<-data.frame(clusteri.d=rep(1:x, each=y),
                  cluster.effect=rep(rnorm(x, 0, 1), each=y),
                  stud.id= c(1:(x*y)),
                  student.effect=rnorm((x*y), 0, sqrt((1-z)/z)))
 data$treat<-sample(1, (x*y), replace=T)

 data2<-data.frame(clusteri.d=rep((x+1):(x*y+x), each=1),
                   cluster.effect=0,
                   stud.id= c((x*y+1):(x*y*2)),
                   student.effect=rnorm((x*y), 0, sqrt((1-z)/z)))

 data2$treat<-sample(0, (x*y), replace=T)
 fdata<-rbind(data,data2)
 }else if(d=="t4"){
   data<-data.frame(clusteri.d=rep(1:x, each=y),
                    cluster.effect=rep(rt(x, 4), each=y),
                    stud.id= c(1:(x*y)),
                    student.effect=rnorm((x*y), 0, sqrt((2-2*z)/z)))
   data$treat<-sample(1, (x*y), replace=T)

   data2<-data.frame(clusteri.d=rep((x+1):(x*y+x), each=1),
```

```
                cluster.effect=0,
                stud.id= c((x*y+1):(x*y*2)),
                student.effect=rnorm((x*y), 0, sqrt((2-2*z)/z)))
  data2$treat<-sample(0, (x*y), replace=T)
  fdata<-rbind(data,data2)

 }else if (d=="t11"){

  data<-data.frame(clusteri.d=rep(1:x, each=y),
                cluster.effect=rep(rt(x, 11), each=y),
                stud.id= c(1:(x*y)),
                student.effect=rnorm((x*y), 0, sqrt(((11/9)-(11/9)*z)/z)))
  data$treat<-sample(1, (x*y), replace=T)

  data2<-data.frame(clusteri.d=rep((x+1):(x*y+x), each=1),
                cluster.effect=0,
                stud.id= c((x*y+1):(x*y*2)),
                student.effect=rnorm((x*y), 0, sqrt(((11/9)-(11/9)*z)/z)))
  data2$treat<-sample(0, (x*y), replace=T)
  fdata<-rbind(data,data2)
 }

 fdata$Yout<-(fdata$treat*fdata$cluster.effect)+fdata$student.effect
fdata
}

model <- function(x) {
 mo1<-lmer(Yout~1+ treat + (0 + treat|clusteri.d), x, REML=FALSE)
 beta0<-summary(mo1)$coefficients[1,1] #####extracting the treatment effect
 sebeta0<-summary(mo1)$coefficients[1,2] #####extracting the standard error of treatment effect
 beta0df<-summary(mo1)$coefficients[1,3] ###### extracting the degrees of freedom
 beta0t<-summary(mo1)$coefficients[1,4]####### extracting the t-value
 pvaluebe0<-summary(mo1)$coefficients[1,5] #####extracting the p-value
 beta<-summary(mo1)$coefficients[2,1] #####extracting the treatment effect
 sebeta<-summary(mo1)$coefficients[2,2] #####extracting the standard error of treatment effect
 betadf<-summary(mo1)$coefficients[2,3] ###### extracting the degrees of freedom
 betat<-summary(mo1)$coefficients[2,4]####### extracting the "t" value
 pvaluebe<-summary(mo1)$coefficients[2,5] #####extracting the p-value
 sig<-summary(mo1)$sigma
 tau<-summary(mo1)[[13]][[1]][[1,1]]
 return(c(beta0, sebeta0,beta0df, beta0t, pvaluebe0, beta, sebeta,betadf, betat, pvaluebe, sig, tau))
}


## (4) design your outer function to do all replications
## it requires an outer function and an inner function


ofu<-function(cell){
 x<-condi[cell,1]
 y<-condi[cell,2]
 z<-condi[cell,3]
 d<-condi[cell,4]
 cell_datasets = replicate(1000, simul(x,y,z,d), simplify = FALSE)

 myrep_list = lapply(cell_datasets, model)
```

```
 myrep = do.call(rbind, myrep_list)
  colnames(myrep) = c("beta0", "sebeta0", "beta0df", "beta0t", "pvaluebe0",
                      "beta","sbeta","betadf", "betat", "pvalue","sig", "tau")
 cell_output = data.frame(myrep, cell = cell)
 savename = paste0("intermediate/Cell", cell, ".Rdata")
 save(cell_output, cell_datasets, file = savename)
 cell_output
}


all_cells = lapply(1:nrow(condi), ofu)
cells_df = rbind_all(all_cells)
write.csv(cells_df,"normalt4t11_1.csv", sep=".\t", row.names=FALSE, col.names=F)
```

Simulation codes for A model: Fixed effects and Power

```
############################################################
##### Codes for simulating data NORMAL DISTRIBUTION ########

##### invoke the packages you need

library(lme4) # Fit the model
library(lmerTest) # estimate the p-values
library(dplyr)

## (1) define your conditions

x<-c(10,30, 50) ### level-2 sample size
y<-c(6, 17, 32) ### within cluster sample size
z<-c(.05,.15,.25) ### intra-class correlation treatment condition
d<-c("Normal", "t4", "t11")


## (2) design your matrix of conditions

condi<- expand.grid(x,y,z,d)
print(condi)

## (3) write your function (inner function): in my case it should be the
## (a) simulation of clusters
## (b) simulation of subjects
## (c) simulation of error distribution
## (d) etc.
set.seed(011179)

simul <-function(x, y, z, d){


 if(d=="Normal"){
 data<-data.frame(clusteri.d=rep(1:x, each=y),
                  cluster.effect=rep(rnorm(x, 0, 1), each=y),
                  stud.id= c(1:(x*y)),
                  student.effect=rnorm((x*y), 0, sqrt((1-z)/z)))
 data$treat<-sample(1, (x*y), replace=T)

 data2<-data.frame(clusteri.d=rep((x+1):(x*y+x), each=1),
```

```
                cluster.effect=0,
                stud.id= c((x*y+1):(x*y*2)),
                student.effect=rnorm((x*y), 0, sqrt((1-z)/z)))

  data2$treat<-sample(0, (x*y), replace=T)
  fdata<-rbind(data,data2)
  }else if(d=="t4"){
    data<-data.frame(clusteri.d=rep(1:x, each=y),
                cluster.effect=rep(rt(x, 4), each=y),
                stud.id= c(1:(x*y)),
                student.effect=rnorm((x*y), 0, sqrt((2-2*z)/z)))
    data$treat<-sample(1, (x*y), replace=T)

    data2<-data.frame(clusteri.d=rep((x+1):(x*y+x), each=1),
                 cluster.effect=0,
                 stud.id= c((x*y+1):(x*y*2)),
                 student.effect=rnorm((x*y), 0, sqrt((2-2*z)/z)))
    data2$treat<-sample(0, (x*y), replace=T)
    fdata<-rbind(data,data2)

  }else if (d=="t11"){

    data<-data.frame(clusteri.d=rep(1:x, each=y),
                cluster.effect=rep(rt(x, 11), each=y),
                stud.id= c(1:(x*y)),
                student.effect=rnorm((x*y), 0, sqrt(((11/9)-(11/9)*z)/z)))
    data$treat<-sample(1, (x*y), replace=T)

    data2<-data.frame(clusteri.d=rep((x+1):(x*y+x), each=1),
                 cluster.effect=0,
                 stud.id= c((x*y+1):(x*y*2)),
                 student.effect=rnorm((x*y), 0, sqrt(((11/9)-(11/9)*z)/z)))
    data2$treat<-sample(0, (x*y), replace=T)
    fdata<-rbind(data,data2)
  }

  fdata$Yout<-(fdata$treat)+(fdata$treat*fdata$cluster.effect)+fdata$student.effect
 fdata
}

model <- function(x) {
  mo1<-lmer(Yout~1+ treat + (0 + treat|clusteri.d), x, REML=FALSE)
  beta0<-summary(mo1)$coefficients[1,1] #####extracting the treatment effect
  sebeta0<-summary(mo1)$coefficients[1,2] #####extracting the standard error of treatment effect
  beta0df<-summary(mo1)$coefficients[1,3] ###### extracting the degrees of freedom
  beta0t<-summary(mo1)$coefficients[1,4]####### extracting the t-value
  pvaluebe0<-summary(mo1)$coefficients[1,5] #####extracting the p-value
  beta<-summary(mo1)$coefficients[2,1] #####extracting the treatment effect
  sebeta<-summary(mo1)$coefficients[2,2] #####extracting the standard error of treatment effect
  betadf<-summary(mo1)$coefficients[2,3] ###### extracting the degrees of freedom
  betat<-summary(mo1)$coefficients[2,4]####### extracting the "t" value
  pvaluebe<-summary(mo1)$coefficients[2,5] #####extracting the p-value
  sig<-summary(mo1)$sigma
  tau<-summary(mo1)[[13]][[1]][[1,1]]
  return(c(beta0, sebeta0,beta0df, beta0t, pvaluebe0, beta, sebeta,betadf, betat, pvaluebe, sig, tau))
}
```

```
## (4) design your outer function to do all replications
## it requires an outer function and an inner function


ofu<-function(cell){
  x<-condi[cell,1]
  y<-condi[cell,2]
  z<-condi[cell,3]
  d<-condi[cell,4]
  cell_datasets = replicate(1000, simul(x,y,z,d), simplify = FALSE)

  myrep_list = lapply(cell_datasets, model)
  myrep = do.call(rbind, myrep_list)

  colnames(myrep) = c("beta0", "sebeta0", "beta0df", "beta0t", "pvaluebe0",
                      "beta","sbeta","betadf", "betat", "pvalue","sig", "tau")

  cell_output = data.frame(myrep, cell = cell)

  savename = paste0("intermediate/Cell", cell, ".Rdata")

  save(cell_output, cell_datasets, file = savename)

  cell_output
}


all_cells = lapply(1:nrow(condi), ofu)
cells_df = rbind_all(all_cells)
write.csv(cells_df,"normalt4t11_1.csv", sep=".\t", row.names=FALSE, col.names=F)
```

Simulation codes for B model: Type I error rate

```
############################################################
##### Codes for simulating data NORMAL DISTRIBUTION ########

##### invoke the packages you need

library(lme4) # Fit the model
library(lmerTest) # estimate the p-values
library(dplyr)

## (1) define your conditions

x<-c(10,30, 50) ### level-2 sample size
y<-c(6, 17, 32) ### within cluster sample size
z<-c(.05,.15,.25) ### intra-class correlation treatment condition
d<-c("Normal", "t4", "t11")


## (2) design your matrix of conditions
```

```
condi<- expand.grid(x,y,z,d)
print(condi)

## (3) write your function (inner function): in my case it should be the
## (a) simulation of clusters
## (b) simulation of subjects
## (c) simulation of error distribution
## (d) etc.
set.seed(011179)

simul <-function(x, y, z, d){


 if(d=="Normal"){
 data<-data.frame(clusteri.d=rep(1:x, each=y),
                  cluster.effect=rep(rnorm(x, 0, 1), each=y),
                  stud.id= c(1:(x*y)),
                  student.effect=rnorm((x*y), 0, sqrt((1-z)/z)),
                  studentX=rnorm((x*y), 0, 1))
 data$treat<-sample(1, (x*y), replace=T)

 data2<-data.frame(clusteri.d=rep((x+1):(x*y+x), each=1),
                   cluster.effect=0,
                   stud.id= c((x*y+1):(x*y*2)),
                   student.effect=rnorm((x*y), 0, sqrt((1-z)/z)),
                   studentX=rnorm((x*y), 0, 1))
 data2$treat<-sample(0, (x*y), replace=T)
 fdata<-rbind(data,data2)


 }else if(d=="t4"){
  data<-data.frame(clusteri.d=rep(1:x, each=y),
                   cluster.effect=rep(rt(x, 4), each=y),
                   stud.id= c(1:(x*y)),
                   student.effect=rnorm((x*y), 0, sqrt((2-2*z)/z)),
                   studentX=rnorm((x*y), 0, 1))
  data$treat<-sample(1, (x*y), replace=T)

  data2<-data.frame(clusteri.d=rep((x+1):(x*y+x), each=1),
                    cluster.effect=0,
                    stud.id= c((x*y+1):(x*y*2)),
                    student.effect=rnorm((x*y), 0, sqrt((2-2*z)/z)),
                    studentX=rnorm((x*y), 0, 1))
  data2$treat<-sample(0, (x*y), replace=T)
  fdata<-rbind(data,data2)



 }else if (d=="t11"){
  data<-data.frame(clusteri.d=rep(1:x, each=y),
                   cluster.effect=rep(rt(x, 11), each=y),
                   stud.id= c(1:(x*y)),
                   student.effect=rnorm((x*y), 0, sqrt(((11/9)-(11/9)*z)/z)),
                   studentX=rnorm((x*y), 0, 1))
  data$treat<-sample(1, (x*y), replace=T)
```

```
   data2<-data.frame(clusteri.d=rep((x+1):(x*y+x), each=1),
                     cluster.effect=0,
                     stud.id= c((x*y+1):(x*y*2)),
                     student.effect=rnorm((x*y), 0, sqrt(((11/9)-(11/9)*z)/z)),
                     studentX=rnorm((x*y), 0, 1))
   data2$treat<-sample(0, (x*y), replace=T)
   fdata<-rbind(data,data2)

  }

  fdata$Yout<-(fdata$treat*fdata$cluster.effect)+fdata$student.effect
 fdata
}

model <- function(x) {
  mo1<-lmer(Yout~1+ treat+ studentX + (0 + treat|clusteri.d), x)
  beta0<-summary(mo1)$coefficients[1,1] #####extracting b0
  sebeta0<-summary(mo1)$coefficients[1,2] #####extracting the standard error of b0
  beta0df<-summary(mo1)$coefficients[1,3] ###### extracting the degrees of freedom
  beta0t<-summary(mo1)$coefficients[1,4]####### extracting the t-value b0
  pvaluebe0<-summary(mo1)$coefficients[1,5] #####extracting the p-value b0
  beta<-summary(mo1)$coefficients[2,1] #####extracting the treatment effect
  sebeta<-summary(mo1)$coefficients[2,2] #####extracting the standard error of treatment effect
  betadf<-summary(mo1)$coefficients[2,3] ###### extracting the degrees of freedom treat eff
  betat<-summary(mo1)$coefficients[2,4]####### extracting the "t" value of treat eff
  pvaluebe<-summary(mo1)$coefficients[2,5] #####extracting the p-value of treat eff
  betaX<-summary(mo1)$coefficient[3,1] ##### extracting X effect
  sebetaX<-summary(mo1)$coefficient[3,2] #### extracting the standard error of X
  betadfX<-summary(mo1)$coefficient[3,3] ### extracting X degrees of freedom
  betatX<-summary(mo1)$coefficient[3,4] ### extracting X t value
  pvalueX<-summary(mo1)$coefficient[3,5] #### extracting the p-value of X
  sig<-summary(mo1)$sigma
  tau<-summary(mo1)[[13]][[1]][[1,1]]
  return(c(beta0, sebeta0,beta0df, beta0t, pvaluebe0, beta, sebeta,betadf, betat, pvaluebe,
         betaX, sebetaX,betadfX,betatX, pvalueX,sig, tau))

}


## (4) design your outer function to do all replications
## it requires an outer function and an inner function


ofu<-function(cell){
  x<-condi[cell,1]
  y<-condi[cell,2]
  z<-condi[cell,3]
  d<-condi[cell,4]
  cell_datasets = replicate(1000, simul(x,y,z,d), simplify = FALSE)

  myrep_list = lapply(cell_datasets, model)
  myrep = do.call(rbind, myrep_list)

  colnames(myrep) = c("beta0", "sebeta0","beta0df", "beta0t", "pvaluebe0",
                      "beta", "sebeta","betadf", "betat", "pvaluebe",
                      "betaX", "sebetaX","betadfX","betatX", "pvalueX","sig", "tau")
```

```
  cell_output = data.frame(myrep, cell = cell)

  savename = paste0("intermediate/Cell", cell, ".Rdata")

  save(cell_output, cell_datasets, file = savename)

  cell_output
}


all_cells = lapply(1:nrow(condi), ofu)
cells_df = rbind_all(all_cells)
write.csv(cells_df,"normalt4t11_1.csv", sep=".\t", row.names=FALSE, col.names=F)
```

Simulation codes for B model: Fixed Effects and Power

```
############################################################
##### Codes for simulating data NORMAL DISTRIBUTION ########

##### invoke the packages you need

library(lme4) # Fit the model
library(lmerTest) # estimate the p-values
library(dplyr)

## (1) define your conditions

x<-c(10,30, 50) ### level-2 sample size
y<-c(6, 17, 32) ### within cluster sample size
z<-c(.05,.15,.25) ### intra-class correlation treatment condition
d<-c("Normal", "t4", "t11")


## (2) design your matrix of conditions

condi<- expand.grid(x,y,z,d)
print(condi)

## (3) write your function (inner function): in my case it should be the
## (a) simulation of clusters
## (b) simulation of subjects
## (c) simulation of error distribution
## (d) etc.
set.seed(011179)

simul <-function(x, y, z, d){


  if(d=="Normal"){
  data<-data.frame(clusteri.d=rep(1:x, each=y),
                   cluster.effect=rep(rnorm(x, 0, 1), each=y),
                   stud.id= c(1:(x*y)),
                   student.effect=rnorm((x*y), 0, sqrt((1-z)/z)),
```

```
                studentX=rnorm((x*y), 0, 1))
  data$treat<-sample(1, (x*y), replace=T)

  data2<-data.frame(clusteri.d=rep((x+1):(x*y+x), each=1),
                    cluster.effect=0,
                    stud.id= c((x*y+1):(x*y*2)),
                    student.effect=rnorm((x*y), 0, sqrt((1-z)/z)),
                    studentX=rnorm((x*y), 0, 1))
  data2$treat<-sample(0, (x*y), replace=T)
  fdata<-rbind(data,data2)


  }else if(d=="t4"){
    data<-data.frame(clusteri.d=rep(1:x, each=y),
                     cluster.effect=rep(rt(x, 4), each=y),
                     stud.id= c(1:(x*y)),
                     student.effect=rnorm((x*y), 0, sqrt((2-2*z)/z)),
                     studentX=rnorm((x*y), 0, 1))
    data$treat<-sample(1, (x*y), replace=T)

    data2<-data.frame(clusteri.d=rep((x+1):(x*y+x), each=1),
                      cluster.effect=0,
                      stud.id= c((x*y+1):(x*y*2)),
                      student.effect=rnorm((x*y), 0, sqrt((2-2*z)/z)),
                      studentX=rnorm((x*y), 0, 1))
    data2$treat<-sample(0, (x*y), replace=T)
    fdata<-rbind(data,data2)



  }else if (d=="t11"){
   data<-data.frame(clusteri.d=rep(1:x, each=y),
                    cluster.effect=rep(rt(x, 11), each=y),
                    stud.id= c(1:(x*y)),
                    student.effect=rnorm((x*y), 0, sqrt(((11/9)-(11/9)*z)/z)),
                    studentX=rnorm((x*y), 0, 1))
    data$treat<-sample(1, (x*y), replace=T)

    data2<-data.frame(clusteri.d=rep((x+1):(x*y+x), each=1),
                      cluster.effect=0,
                      stud.id= c((x*y+1):(x*y*2)),
                      student.effect=rnorm((x*y), 0, sqrt(((11/9)-(11/9)*z)/z)),
                      studentX=rnorm((x*y), 0, 1))
    data2$treat<-sample(0, (x*y), replace=T)
    fdata<-rbind(data,data2)

  }

  fdata$Yout<-(fdata$treat)+(fdata$treat*fdata$cluster.effect)+fdata$studentX+fdata$student.effect
 fdata
}

model <- function(x) {
  mo1<-lmer(Yout~1+ treat+ studentX + (0 + treat|clusteri.d), x, REML=FALSE)
  beta0<-summary(mo1)$coefficients[1,1] #####extracting b0
  sebeta0<-summary(mo1)$coefficients[1,2] #####extracting the standard error of b0
```

```
  beta0df<-summary(mo1)$coefficients[1,3] ###### extracting the degrees of freedom
  beta0t<-summary(mo1)$coefficients[1,4]####### extracting the t-value b0
  pvaluebe0<-summary(mo1)$coefficients[1,5] #####extracting the p-value b0
  beta<-summary(mo1)$coefficients[2,1] #####extracting the treatment effect
  sebeta<-summary(mo1)$coefficients[2,2] #####extracting the standard error of treatment effect
  betadf<-summary(mo1)$coefficients[2,3] ###### extracting the degrees of freedom treat eff
  betat<-summary(mo1)$coefficients[2,4]####### extracting the "t" value of treat eff
  pvaluebe<-summary(mo1)$coefficients[2,5] #####extracting the p-value of treat eff
  betaX<-summary(mo1)$coefficient[3,1] ##### extracting X effect
  sebetaX<-summary(mo1)$coefficient[3,2] #### extracting the standard error of X
  betadfX<-summary(mo1)$coefficient[3,3] ### extracting X degrees of freedom
  betatX<-summary(mo1)$coefficient[3,4] ### extracting X t value
  pvalueX<-summary(mo1)$coefficient[3,5] #### extracting the p-value of X
  sig<-summary(mo1)$sigma
  tau<-summary(mo1)[[13]][[1]][[1,1]]
  return(c(beta0, sebeta0,beta0df, beta0t, pvaluebe0, beta, sebeta,betadf, betat, pvaluebe,
         betaX, sebetaX,betadfX,betatX, pvalueX,sig, tau))

}


## (4) design your outer function to do all replications
## it requires an outer function and an inner function


ofu<-function(cell){
  x<-condi[cell,1]
  y<-condi[cell,2]
  z<-condi[cell,3]
  d<-condi[cell,4]
  cell_datasets = replicate(1000, simul(x,y,z,d), simplify = FALSE)

  myrep_list = lapply(cell_datasets, model)
  myrep = do.call(rbind, myrep_list)

  colnames(myrep) = c("beta0", "sebeta0","beta0df", "beta0t", "pvaluebe0",
                      "beta", "sebeta","betadf", "betat", "pvaluebe",
                      "betaX", "sebetaX","betadfX","betatX", "pvalueX","sig", "tau")

  cell_output = data.frame(myrep, cell = cell)

  savename = paste0("intermediate/Cell", cell, ".Rdata")

  save(cell_output, cell_datasets, file = savename)

  cell_output
}


all_cells = lapply(1:nrow(condi), ofu)
cells_df = rbind_all(all_cells)
write.csv(cells_df,"normalt4t11_1.csv", sep=".\t", row.names=FALSE, col.names=F)
```